Université Lille1 Sciences et Technologies

Numéro d'ordre:
41090

Université des Sciences et Technologies de Lille
École Doctorale Sciences Pour l'Ingénieur Lille Nord-de-France

T H È S E

pour obtenir le titre de

Docteur en Sciences
Mention : MATHÉMATIQUES APPLIQUÉES
Présentée et soutenue le 10 Juin 2013

Analyse de la convergence de l'algorithme FastICA: Echantillon de taille finie et infinie

par

Tianwen Wei

Composition du Jury:

Directeur de thèse :	Azzouz Dermoune	-	Université Lille 1
Rapporteurs :	Pierre Comon	-	Université de Grenoble
	Ali Mohammad-Djafari	-	Université Paris-Sud
Examinateurs :	Stephane Gaiffas	-	Ecole Polytechnique Paris
	Guillaume Lecue	-	Université Paris-Est
	Cristian Preda	-	Université Lille 1
	Nicolas Wicker	-	Université Lille 1

Remerciements

Je souhaite tout d'abord exprimer ma profonde gratitude au Professeur Azzouz Dermoune, mon directeur de thèse pour avoir dirigé ce travail. Sa rigueur, sa clairvoyance, sa patience ainsi que le soutien qu'il m'a toujours apporté, m'ont permis de mener à bien cette thèse. Je n'oublierai jamais ses qualités scientifiques et humaines qui ont contribué énormément à la progression de mes travaux de recherche.

Mes sincère remerciements sont également adressés à Nadji Rahmania, professeur et collaborateur dans notre groupe de travail, pour ses conseils précieux qu'il m'a accordé tout au long de ces années de travail.

Un grand merci à Pierre Comon et Ali-Mohammad Djafari, qui ont accepté de rapporter cette thèse. Leurs lectures attentives m'ont permis d'améliorer mon travail. Je tiens à remercier également les autres membres du jury: Guillaume Lecue, Stephane Gaiffas, Cristian Preda et Nicolas Wicker, qui sont venus de loin ou de près.

Je suis très honoré que Appo Hyvärinen, le créateur de l'algorithme FastICA, qui est l'un des spécialistes internationaux les plus reconnus de mon domaine de recherche, ait accepté de nous acceuillir à l'université d'Helsinki. Je le remercie vivement.

J'exprime toute ma reconnaissance à mes anciens enseignants et plus particulièrement à Antoine Ayache, Tran Viet Chi, Youri Davydov et Radu Stoica, qui par la qualité de leurs cours, ont considérablement contribué à me donner envie de faire une thèse. Je suis aussi reconnaissant à Thierry Goudon, pour l'aide qu'il m'a apporté quand je faisait mon masteur.

Je tiens à remercier les anciens doctorants chinois du bureau, Qidi Peng et Ying Chen, pour les moments merveilleurs que nous avions passé ensemble pendant ces années. Je n'ai pas oublié non plus ma chère Ying, ainsi que les autres amis chinois et internationaux sur Lille, Jianwei, Cheng, Xian, Jing, Chen, Zuqi, Martin, Elsa, Sophie, Safa, Vincent, Xavier, etc.

J'exprime ma profonde gratitude à Changgui Zhang. C'est grâce à lui que j'aie l'opportunié de venir étudier en France, ce qui a radicalement changé ma vie.

Finalement, je pense beaucoup à ma mère Jie Yang et mon père Yueqing Wei, qui m'ont soutenu unconditionnellement malgré la distance tout au long de mes études en France. Je les remercie du fond de mon coeur pour leur sacrifice énorme.

Analyse de la convergence de l'algorithme FastICA: Echantillon de taille finie et infinie

Résumé

L'algorithme FastICA est l'un des algorithmes les plus populaires dans le domaine de l'analyse en composantes indépendantes (ICA). Il existe deux versions de FastICA: Celle qui correspond au cas où l'échantillon est de taille infinie, et celle qui traite de la situation concrète, où seul un échantillon de taille finie est disponible. Dans cette thèse, nous avons fait une étude détaillée des vitesses de convergence de l'algorithme FastICA dans le cas où la taille de l'échantillon est finie ou infinie, et nous avons établi cinq critères pour le choix des fonctions de non-linéarité.

Dans les trois premiers chapitres, nous avons introduit le problème de l'ICA et revisité les résultats existants. Dans le Chapitre 4, nous avons étudié la convergence du FastICA empirique et le lien entre la limite de FastICA empirique et les points critiques de la fonction de contraste empirique. Dans le Chapitre 5, nous avons utilisé la technique du M-estimateur pour obtenir la normalité asymptotique et la matrice de covariance asymptotique de l'estimateur FastICA. Ceci nous a permis aussi de déduire quatre critères pour choisir les fonctions de non-linéarité. Un cinquième critère de choix de non-linéarité a été étudié dans le chapitre 6. Ce critère est basé sur une étude fine de la vitesse de convergence de FastICA empirique. Nous avons illustré chaque chapitre par des résultats numériques qui valident nos résultats théoriques.

Contents

List of Notations

Typesetting convention

a, b, c	lower case letter signifies real scalar
$\mathbf{a}, \mathbf{b}, \mathbf{c}$	boldface lower case letter signifies column real vector
$\mathbf{A}, \mathbf{B}, \mathbf{C}$	boldface upper case letter signifies real matrix
$(\mathbf{A})_{ij}$	(i, j)-th entry of the matrix $\mathbf{A}$
$\mathbf{x}(t)$	the t-th realization of random vector $\mathbf{x}$

Operations

$\mathbf{a}^\mathsf{T}$, $\mathbf{A}^\mathsf{T}$	the transpose of a vector or matrix
$\mathbb{E}[\cdot]$	the mathematical expectation operator
$\mathrm{Cov}(\mathbf{x})$	the variance or covariance matrix of $\mathbf{x}$
$\mathrm{vec}(\cdot)$	the operation that reshapes the columns of a matrix into a long column vector
$\|\cdot\|$	Euclidean norm for vector; induced L_2 norm (spectral norm) for matrix
$\stackrel{\text{def}}{=}$	be defined as

Particular notations

$\mathbf{A}$	the mixing matrix
$\mathbf{I}$	the identity matrix
$\Pi_{\mathbf{v}}$	the orthogonal projection matrix that project onto span($\mathbf{v}$)
$\Pi_{\mathbf{v}}^{\perp}$	the orthogonal projection matrix that project onto span($\mathbf{v}$)$^{\perp}$
$\mathbf{a}$	a generic column of the mixing matrix $\mathbf{A}$
$\mathbf{a}_i$	the i-th column of the mixing matrix $\mathbf{A}$
$\mathbf{e}_i$	the i-th column of the identity matrix
$\mathbf{s} = (s_1, \ldots, s_d)^{\mathsf{T}}$	the source signal
$\mathbf{x} = (x_1, \ldots, x_d)^{\mathsf{T}}$	the observed signal
$\mathbb{R}^d$	the set of d-dimensional real column vectors
$\mathbb{R}^{n \times m}$	the set of $n \times m$ real matrices
$\mathcal{S}$	the set of vectors having unit norm
$\mathscr{B}_r(\mathbf{v})$	the set $\{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w} - \mathbf{v}\| = r\}$
span($\mathbf{v}$)	the linear subspace spanned by vector $\mathbf{v}$
$G(\cdot)$	the nonlinearity function
$G(\cdot, \mu)$	the theoretical contrast function with underlying nonlinearity G
$G(\cdot, \mu_N^k)$	the empirical contrast function with underlying nonlinearity G
$g(\cdot)$	the derivative of nonlinearity $G(\cdot)$
$\mathcal{I}(\mathbf{z})$	the mutual information of random vector $\mathbf{z}$
$\mathcal{J}(\mathbf{z})$	the negentropy of random vector z
KL(p||q)	the Kullback-Leibler divergence between pdf p and q
$\boldsymbol{f}(\cdot, \mu)$	the theoretical FastICA function
$\boldsymbol{f}(\cdot, \mu_N^k)$	the empirical FastICA function
$\mathcal{L}(\cdot)$	the Lagrange function
μ	the probability distribution of the observed signal $\mathbf{x}$
μ_N^k	the k-th discrete probability distribution based on a sample of $\mathbf{x}$

Chapter 1
Introduction

The author's work during four years of study consists of two independent parts. The first part concerns the study of the generalized linear model, which leads to the publication (Dermoune, Rahmania, & Wei, 2012). Due to the lack of time, this work will not be incorporated in this thesis. The second part of the author's work concerns the study of the FastICA algorithm. The main results of this part is published (Dermoune & Wei, 2013) in the journal *IEEE Transaction on Signal Processing*. This thesis is a completion of this paper.

The *Blind Source Separation* (Comon & Jutten, 2010; Jutten & Comon, n.d.; Jutten & Taleb, 2000), often abbreviated as BSS, is a statistical and computational method for revealing hidden factors that underlie sets of random variable or signals. The term "blind" is intended to imply that such methods can separate data into source signals with the absence or prior information about the nature of the source signals and the process that mixes these signals. The BSS problem, first formulated in the early 80s, is a fast growing research area in the past thirty years. It has drawn great attention of many researchers, notably those from neural network and signal processing community, and it has become a widely used data analysis and signal processing technique with application in many diverse fields, such as biomedical signal processing, image processing, acoustic signal separation, telecommunications, fault diagnosis, and financial time series (Comon & Jutten, 2010; Hyvärinen, Karhunen, & Oja, 2001; Makeig, Bell, Jung, & Sejnowski, 1996; M. Ichir, 2006; Vigario & Oja, 2008; Makino, Lee, & Sawada, 2007; Brandstein & Ward, 2001).

A general framework for solving BSS problems is called the *Independent Component Analysis* (ICA) (Hyvärinen et al., 2001; Stone, 2004; Jutten, 1987; Comon, 1994; Hyvärinen & Oja, 2000), which is based on, as the name suggests, the simple and fundamental assumption that the unknown sources are statistically independent. This assumption is physically realistic due to the fact that different physical processes (e.g. different people speaking) generate statistically independent signals. Aside from the independence of source signals, typical ICA assumptions also include the linearity and the instantaneousness of the mixture. Even though there exist some methods for which an algebraic solution to the ICA problem may be found, other iterative methods are very popular. Particularly, in practical real-world problems, people work with a large number of observed variables and data points, in which case an efficient numerical algorithm is even necessary, since

the precise algebraic solution can only be estimated from the data. Up to date, there exist various algorithms (Delfosse & Loubaton, 1995; Cardoso & Souloumiac, 1993; Tugnait, 1997; Comon & Moreau, 1997; Chevalier, Albera, Comon, & Ferreol, 2004; Zarzoso & Comon, 2010) in the domain of ICA, among which the so called *FastICA* algorithm, proposed by Hyvärinen and Oja (Hyvärinen & Oja, 2000, 1997) from Finnish school, is arguably the most popular one. The success of FastICA can be attributed its simplicity, ease of implementation and flexibility to choose the nonlinearity function.

There are two versions of FastICA: the theoretical FastICA and the empirical FastICA. The former corresponds to the ideal case that the mathematical expectation appeared in the formulation of the algorithm can be precisely computed, while the latter deals with the practical situation, where only a finite sample is available and hence the mathematical expectation must be approximated by a sample average. The theoretical FastICA has already been extensively studied by many researchers during the past decade (Hyvärinen & Oja, 1997; Hyvärinen, 1999; Regalia & Kofidis, 2003; Oja, 2002; Douglas, 2003), while the empirical FastICA still poses many theoretical and numerical problems.

In this thesis, we are particularly interested in the following questions:

1) Does the empirical FastICA algorithm always converge?

2) The empirical FastICA is an estimator of the theoretical FastICA. What about its asymptotic performance?

3) Does the empirical FastICA algorithm have a quadratic convergence speed? What is the best choice of the nonlinearity function in terms of convergence speed?

Even though there exist some partial answers to these questions in the literature (Hyvärinen, 1997; Tichavsky, Koldovsky, & Oja, 2006; Oja & Yuan, 2006; Ollila, 2010; Leshem & van der Veen, 2008), most of them are based on simulations; formal developments are often lacking. This thesis aims at filling this lack as well as providing some insight, too.

This thesis is organized as follows. Chapter 2 is preliminary. In this chapter, we will introduce the data model and the assumptions, the notion of contrast function, preprocessing procedure, and finally the FastICA algorithm. Experienced readers may skip this chapter. In Chapter 3, we will develop a new method to reestablish all the classical results concerning the theoretical FastICA, such as the quadratic convergence speed and the limit of FastICA as the local minimizer of the contrast function. In Chapter 4, we will use the same method to tackle the empirical FastICA and give answer to Question 1) listed above. We will propose four empirical FastICA estimators, defined as the limit of the empirical FastICA algorithm, each with respect to a particular measure based on the sample. We will show that with

probability one, all of the four estimators are well defined in the sense that the respective algorithm is convergent. Chapter 5 is devoted to the study of the asymptotic performance of empirical FastICA estimator. We will use the technique of M-estimator to derive the asymptotic normality of empirical FastICA estimator and its asymptotic covariance matrix. Besides, we will compare four criteria which measure the asymptotic performance of the empirical FastICA estimator. The main result of this chapter is Theorem 5.1.1, which answers Question 2). Finally, we adress Question 3) in Chapter 6. We will present a new criterion for the nonlinearity function based upon the convergence speed of FastICA, and give some numerical results.

Chapter 2

Preliminaries

Contents

In this chapter, we will introduce briefly the data model of ICA, data centering and whitening, the important notions of contrast functions and eventually an iterative method called the FastICA algorithm. We will distinguish from the very beginning two versions of ICA that differ in nature: the theoretical ICA, where we work with a theoretical framework (i.e. the exact probability distribution of the observed signal is supposed to be known) and do not care its real-world realization. In this case, the mathematical expectation is calculable, hence everything under consideration is purely deterministic. Next, we will consider the empirical ICA, which corresponds to the practical situation, where we do not have a direct access of the distribution of the observed signal and have to estimate it through sampling.

2.1 Theoretical ICA

2.1.1 Theoretical ICA Model

Let us start by introducing the noiseless linear ICA model:

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{2.1.1}$$

where

- $\mathbf{s} \overset{\text{def}}{=} (s_1, \ldots, s_d)^\mathsf{T}$ denotes the unknown *source signals*. The components $s_1, \ldots, s_d$ are mutually independent and none of them is Gaussian.
- $\mathbf{x} \overset{\text{def}}{=} (x_1, \ldots, x_d)^\mathsf{T}$ denotes the *observed signals*.

- The unknown mixing matrix $\mathbf{A}$ is a non-singular square matrix.

In the sequel, we will use the Greek letter μ to denote the probability distribution of the observed signal $\mathbf{x}$. ICA model 2.1.1 will be referred to as the **theoretical ICA model**, which means that there is no sampling involved and the signal $\mathbf{x}$ is perfectly observed in the sense that its exact probability distribution μ is known. The hypothesis that the source signal $\mathbf{s}$ has independent components is fundamental for ICA, while the non-gaussianity of $\mathbf{s}$ is necessary[1] for the separation of the sources (Comon, 1994). Besides, matrix $\mathbf{A}$ being square means $\dim(\mathbf{x}) = \dim(\mathbf{s})$, which is not very restrictive. Clearly, if $\mathbf{A}$ is not square, it is not invertible, but if $\dim(\mathbf{x}) > \dim(\mathbf{s}) = \operatorname{rank}(\mathbf{A}) = d$, it is still possible to recover the sources. In this so called *overdetermined* case, it suffices to discard some components of the mixture vector $\mathbf{x}$ that can be generated by a linear combination of other rows of the mixing matrix. One can use Principal Component Analysis (PCA), to project the mixture data to a d-dimensional space without any loss of information, see (Hyvärinen et al., 2001) for more detail. Although other models (noisy, nonlinear or with convolutive mixture) are also considered by some authors, model (2.1.1) with the assumptions given above is the simplest but also the most common one for ICA (Hyvärinen et al., 2001; Hyvärinen & Oja, 2000).

The aim of ICA is to recover the independent components of the source signal $\mathbf{s}$ based on the knowledge of μ only. It's beneficial to recall here the notion of independence for a family of random variables:

Definition 2.1.1 (Independence)**.** *Let $z_1, \ldots, z_d$ be random variables having probability density function $p_{z_1}, \ldots, p_{z_d}$. We say $z_1, \ldots, z_d$ are mutually independent if and only if their joint probability density function $p_{\mathbf{z}}$ satisfies*

$$p_{\mathbf{z}} = \prod_{i=1}^{d} p_{z_i}.$$

The recovery of $\mathbf{s}$ can be achieved by finding the inverse of the mixing matrix $\mathbf{A}$. However, under current assumptions the inverse $\mathbf{A}^{-1}$ cannot be identified. One reason is that with both $\mathbf{A}$ and $\mathbf{s}$ being unknown to us, we can never determine the scale (i.e. variance) of $\mathbf{s}$ by the knowledge of $\mathbf{x}$ alone. This is because any scalar multiplier to components of $\mathbf{s}$ could always be canceled by multiplying $\mathbf{A}$ by a diagonal matrix. See the following example:

Example 2.1.2. Let ξ_1, ξ_2 be two random variables whose distributions are not Gaussian. Consider the following two ICA models:

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \qquad \mathbf{x}' = \mathbf{A}'\mathbf{s}',$$

[1] To be precise, at most one component of $\mathbf{s}$ can be Gaussian. In this thesis, we suppose that none of the sources can be Gaussian for simplicity.

where $\mathbf{A} = \mathbf{I}$, $\mathbf{s} = (\xi_1, \xi_2)^\mathsf{T}$, $\mathbf{A}' = \mathrm{diag}\{2, 2\}$ and $\mathbf{s}' = (\frac{\xi_1}{2}, \frac{\xi_2}{2})^\mathsf{T}$. Clearly, we have $\mathbf{x} = \mathbf{x}'$ but $\mathbf{s} \neq \mathbf{s}'$. Based solely upon the knowledge of the observed signal, one cannot determine the source signal nor the mixing matrix, .

This inherent indeterminacy of ICA model 2.1.1 can be reduced by simply making the convention $\mathrm{Cov}(\mathbf{s}) = \mathbf{I}$. This is what we are going to do next. In the sequel, the following hypothesis is always assumed:

Assumption 1. *The components of* $\mathbf{s}$ *have unit variance, i.e.* $\mathrm{Cov}(\mathbf{s}) = \mathbf{I}$.

Assumption 1 still cannot guarantee the identifiability of $\mathbf{A}^{-1}$. However, assuming $\mathrm{Cov}(\mathbf{s}) = \mathbf{I}$ enables us to recover $\mathbf{s}$ up to the sign and a permutation. More precisely, if we are able to find a matrix $\mathbf{W}$, such that the components of $\mathbf{z} \overset{\text{def}}{=} \mathbf{W}\mathbf{x}$ are mutually independent and have unit variance, then we must have $\mathbf{z} = \mathbf{\Lambda}\mathbf{P}\mathbf{s}$ where $\mathbf{P}$ is a permutation matrix and $\mathbf{\Lambda}$ is a diagonal matrix satisfying $\mathbf{\Lambda}^2 = \mathbf{I}$. Before proving this result, let us first give the matrix $\mathbf{W}$ a name:

Definition 2.1.3. *If a matrix* $\mathbf{W}$ *is such that* $\mathbf{z} \overset{\text{def}}{=} \mathbf{W}\mathbf{x}$ *has mutually independent component with unit variance, then* $\mathbf{W}$ *is called a* **demixing matrix**. *A row of a demixing matrix will be called a* **demixing vector**.

This following result can be find in (Comon, 1994).

Theorem 2.1.4. *Let* $\mathbf{s}$ *be a vector of independent components, of which at most one is Gaussian, and whose densities are not reduced to a point-like mass. Let* $\mathbf{B}$ *be an orthogonal matrix in* $\mathbb{R}^{d\times d}$ *and* $\mathbf{z} = (z_1, \ldots, z_d)$ *the vector* $\mathbf{z} = \mathbf{B}\mathbf{s}$. *Then the following three properties are equivalent:*

(i) The components z_i *are pairwise independent.*

(ii) The components z_i *are mutually independent.*

(iii) $\mathbf{B} = \mathbf{\Lambda}\mathbf{P}$ *where* $\mathbf{\Lambda}$ *is diagonal and* $\mathbf{P}$ *is a permutation.*

Theorem 2.1.4 indicates that if $\mathbf{z} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{s}$ has mutually independent components, and if both $\mathbf{z}$ and $\mathbf{s}$ have unit variance, then $\mathbf{B} = \mathbf{W}\mathbf{A} = \mathbf{\Lambda}\mathbf{P}$ with $\mathbf{\Lambda}^2 = \mathbf{I}$. To see this, we note first that $\mathrm{Cov}(\mathbf{z}) = \mathrm{Cov}(\mathbf{\Lambda}\mathbf{P}\mathbf{s}) = \mathbf{I}$, or $\mathbf{\Lambda}\mathbf{P}\mathbf{P}^\mathsf{T}\mathbf{\Lambda}^\mathsf{T} = \mathbf{I}$. Besides, since permutation matrices are orthogonal matrices, we have $\mathbf{P}\mathbf{P}^\mathsf{T} = \mathbf{I}$. Then it follows that $\mathbf{\Lambda}^2 = \mathbf{I}$, i.e. the diagonal elements of $\mathbf{\Lambda}$ are ± 1. In view of this discussion, we deduce the following result:

Corollary 2.1.5. *A matrix* $\mathbf{W}$ *is a demixing matrix if and only if there exists a permutation matrix* $\mathbf{P}$ *and a diagonal matrix* $\mathbf{\Lambda}$ *with* $\mathbf{\Lambda}^2 = \mathbf{I}$, *such that* $\mathbf{W} = \mathbf{\Lambda}\mathbf{P}\mathbf{A}^{-1}$.

Finally, we are ready to formally state the ICA problem.

Problem 2.1.6 (Theoretical ICA problem)**.** *The theoretical ICA problem consists of finding a demixing matrix* $\mathbf{W}$ *based solely upon the distribution* μ *of the observed signal* $\mathbf{x}$.

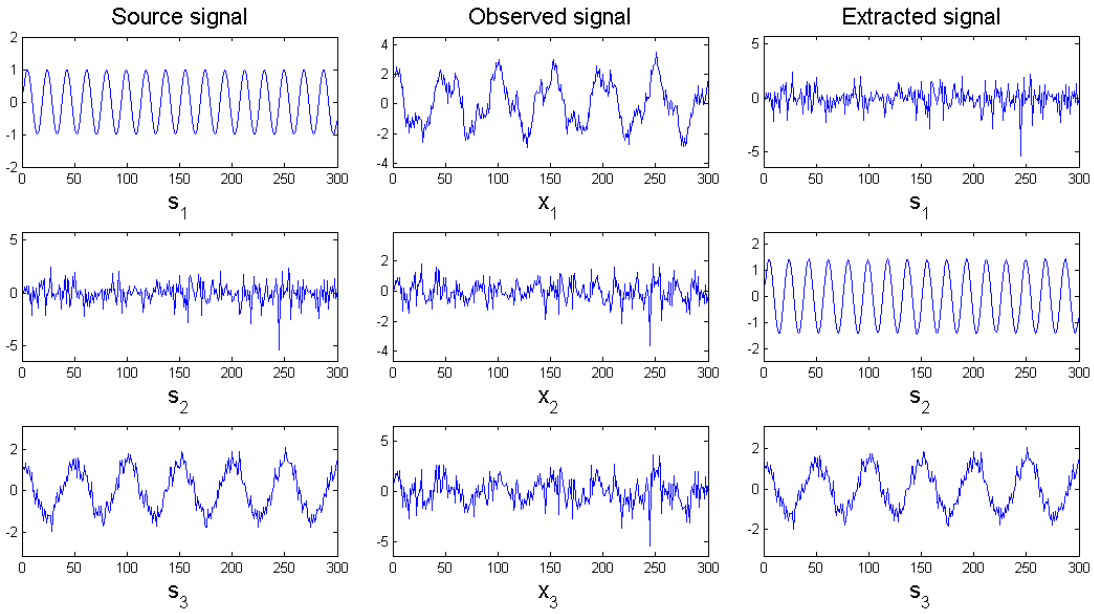

Figure 2.1: Recovering 3 independent components from their mixtures using ICA.

2.1.2 Data preprocessing

Before implementing any ICA method to solve ICA problem 2.1.6, it is usually convenient (necessary in the case of FastICA) to preprocess the data, so that it would be as simple as possible to deal with. Common preprocessing procedures originate from *Principle Component Analysis* (PCA), including *centering* and *whitening*.

Centering is always the first preprocessing procedure. It consists of subtracting from $\mathbf{x}$ its mean $\mathbb{E}[\mathbf{x}]$ to fabricate a new random vector that has zero mean. Whitening normally comes after centering. It aims at transforming the centered signal into a white one, i.e. the one whose components having unit variance and decorrelated. This can be achieved by multiplying the centered signal by $\mathrm{Cov}(\mathbf{x})^{-\frac{1}{2}}$. For model 2.1.1, centering and whitening are feasible since both $\mathbb{E}[\mathbf{x}]$ and $\mathrm{Cov}(\mathbf{x})$ can be drawn from μ.

Definition 2.1.7. *For theoretical ICA model, the preprocessed signal is defined as*

$$\tilde{\mathbf{x}} \overset{\text{def}}{=} \mathrm{Cov}(\mathbf{x})^{-\frac{1}{2}}(\mathbf{x} - \mathbb{E}[\mathbf{x}]). \tag{2.1.2}$$

Now let's look closely at (2.1.2). We have

$$\tilde{\mathbf{x}} = \mathrm{Cov}(\mathbf{x})^{-\frac{1}{2}}\mathbf{A}\big(\mathbf{s} - \mathbb{E}[\mathbf{s}]\big) = \widetilde{\mathbf{A}}\tilde{\mathbf{s}}, \tag{2.1.3}$$

where $\widetilde{\mathbf{A}} \overset{\text{def}}{=} \mathrm{Cov}(\mathbf{x})^{-\frac{1}{2}}\mathbf{A}$ and $\tilde{\mathbf{s}} \overset{\text{def}}{=} \mathbf{s} - \mathbb{E}[\mathbf{s}]$. Clearly, both $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{s}}$ have zero mean and independent components. Besides, we claim that the matrix $\widetilde{\mathbf{A}}$ is a

orthogonal. In fact, by the assumption $\mathrm{Cov}(\mathbf{s}) = \mathbf{I}$, we have $\mathrm{Cov}(\mathbf{x}) = \mathbf{A}\mathbf{A}^\mathsf{T}$. It follows that

$$\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\mathsf{T} = (\mathbf{A}\mathbf{A}^\mathsf{T})^{-\frac{1}{2}}\mathbf{A}\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T})^{-\frac{1}{2}} = \mathbf{I}.$$

In view the discussion above, we can deduce the following result:

Lemma 2.1.8. *Through centering and whitening, we can always transform the theoretical ICA model into an equivalent one, namely*

$$\tilde{\mathbf{x}} = \tilde{\mathbf{A}}\tilde{\mathbf{s}} \tag{2.1.4}$$

where $\widetilde{\mathbf{A}} \overset{\text{def}}{=} (\mathbf{A}\mathbf{A}^\mathsf{T})^{-1/2}\mathbf{A}$ *is orthogonal and* $\tilde{\mathbf{s}} \overset{\text{def}}{=} \mathbf{s} - \mathbb{E}[\mathbf{s}]$ *has zero mean.*

Now that we can always work with model 2.1.4, the following additional assumptions can be added for theoretical ICA model 2.1.1 without loss of generality :

Assumption 2. *1) The source signal* $\mathbf{s}$ *has zero mean, i.e.* $\mathbb{E}[\mathbf{s}] = 0$,

2) The mixing matrix $\mathbf{A}$ *is orthogonal, i.e.* $\mathbf{A}\mathbf{A}^\mathsf{T} = \mathbf{I}$.

Assumption 2 together with Corollary 2.1.5 lead us to the following result:

Proposition 2.1.9. *The demixing matrix* $\mathbf{W}$ *is orthogonal. Moreover, there exists a permutation matrix* $\mathbf{P}$ *and a diagonal matrix* $\mathbf{\Lambda}$ *with* $\mathbf{\Lambda}^2 = \mathbf{I}$, *such that* $\mathbf{W} = \mathbf{\Lambda}\mathbf{P}\mathbf{A}^\mathsf{T}$.

Proposition 2.1.9 tells us that the rows of the demixing matrix $\mathbf{W}$ are $\mathbf{a}_1^\mathsf{T}, \ldots, \mathbf{a}_d^\mathsf{T}$ up to the sign, where $\mathbf{a}_1, \ldots, \mathbf{a}_d$ are columns of $\mathbf{A}$.

2.1.3 Contrast function

The demixing matrix $\mathbf{W}$ can be obtained by optimizing a criterion (Comon & Jutten, 2010; Comon, 1994; Vrins, 2007) called *contrast or contrast function* that measures the dependence between the components of $\mathbf{W}\mathbf{x}$. One traditional measure of dependence, widely used in the community of ICA is the *Kullback-Leibler divergence*, also known as the *relative entropy*. In statistics, it belongs to a large class called *f-divergence*, which can be considered as a kind of "distance" between two probability distributions. In the sequel, we consider those distributions having a probability density function for simplicity.

Example 2.1.10 (Kullback-Leibler divergence). Let p, q be two d-variate density functions, with p being absolutely continuous with respect to q. Then the KL divergence between p and q is defined as:

$$\mathrm{KL}(p||q) \overset{\text{def}}{=} \int_{\mathbb{R}^d} p(\mathbf{t}) \log \frac{p(\mathbf{t})}{q(\mathbf{t})} \mathrm{d}\mathbf{t} = \mathbb{E}\Big[\log \frac{p(\mathbf{u})}{q(\mathbf{u})}\Big], \tag{2.1.5}$$

where $\mathbf{u}$ is a random vector having probability density function p.

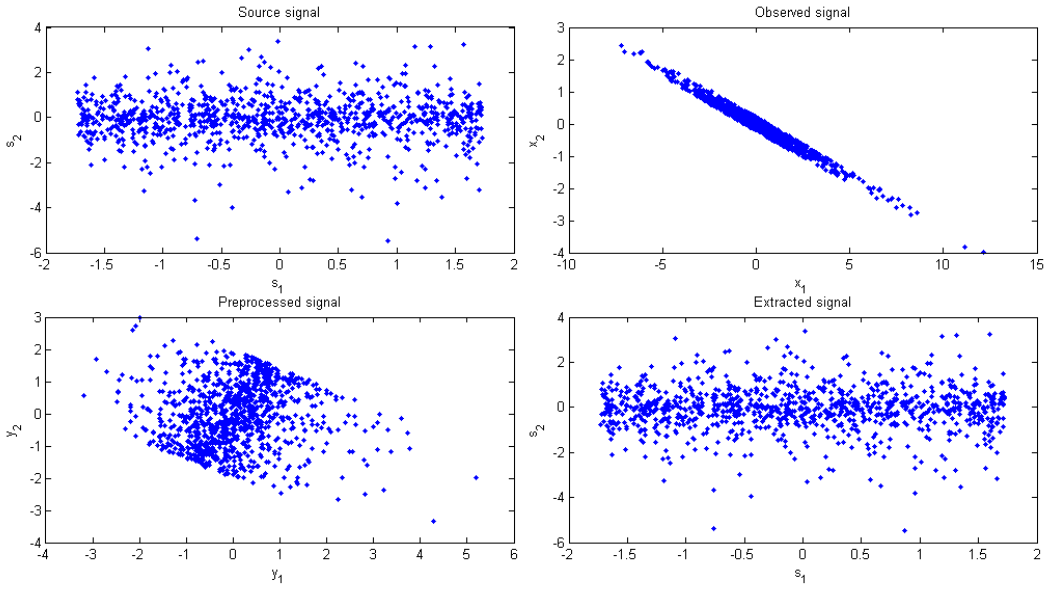

Figure 2.2: Comparison between source signal, observed signal, preprocessed signal and extracted signal in a 2d plane.

Kullback-Leibler divergence is not a true metric distance, since it is not symmetrical, i.e. $\mathrm{KL}(p||q) \neq \mathrm{KL}(q||p)$, and does not satisfy the triangle inequality. Nevertheless, we have the property $\mathrm{KL}(p||q) \geq 0$ with equality if and only if $p = q$ almost everywhere. To see this, we define $Y \stackrel{\mathrm{def}}{=} q(\mathbf{u})/p(\mathbf{u})$, then there holds $\mathbb{E}[Y] = 1$. Now applying Jensens inequality, we get:

$$\mathrm{KL}(p||q) = -\mathbb{E}\Big[\log \frac{q(\mathbf{u})}{p(\mathbf{u})}\Big] = -\mathbb{E}\Big[\log Y\Big] \geq \log \mathbb{E}[Y] = 0.$$

This important property ensures that Kullback-Leibler divergence is a legitimate measure of distance between two probability densities.

In the context of ICA, we are interested in the divergence between the joint density and the product of the marginal densities of a random vector. This thought leads us to the notion of *mutual information*.

Definition 2.1.11 (Mutual information)**.** *The mutual information of a random vector* $\mathbf{z} = (z_1, \ldots, z_d)^\mathsf{T}$ *is defined as* $\mathcal{I}(\mathbf{z}) \stackrel{\mathrm{def}}{=} \mathrm{KL}\Big(p_\mathbf{z} \Big|\Big| \prod_{k=1}^{d} p_{z_k}\Big)$.

The notion of mutual information originates from the information theory. It can be interpreted as the code length reduction obtained by coding the whole vector instead of the separate components. In general, better codes can be obtained by coding the whole vector. However, if the components are independent, they give no information on each other, and one could just as well code the variables separately without increasing code length. In

fact, by the property of Kullback-Leibler divergence and Definition 2.1.1, we have $\mathcal{I}(\mathbf{z}) \geq 0$ and $\mathcal{I}(\mathbf{z}) = 0$ if and only if the components of $\mathbf{z}$ are mutually independent. Therefore, for our ICA model 2.1.1, if we can find a matrix $\mathbf{W}^*$ such that $\mathcal{I}(\mathbf{W}^*\mathbf{x}) = \mathcal{I}(\mathbf{z}^*) = 0$, then our ICA problem is solved. Note that the mutual information is always non negative, hence $\mathbf{W}^*$ can be obtained by solving the following optimization problem:

$$\min_{\mathbf{W}} \mathcal{I}(\mathbf{W}\mathbf{x}) \text{ subject to } \mathbf{W} \text{ being non singular.} \tag{2.1.6}$$

This problem is difficult to tackle in its original form, but it can be effectively simplified by the preprocessing procedure introduced in Section 2.1.2. It can be shown (Comon, 1994) that

$$\mathcal{I}(\mathbf{z}) = \mathcal{J}(\mathbf{z}) - \sum_{k=1}^{d} \mathcal{J}(z_k) + \frac{1}{2} \log \frac{\prod_{k=1}^{d} \mathrm{Cov}(z_k)}{\det(\mathrm{Cov}(\mathbf{z}))}, \tag{2.1.7}$$

where $\mathcal{J}(\mathbf{z}) \stackrel{\text{def}}{=} \mathrm{KL}(p_{\mathbf{z}} || \phi_{\mathbf{z}})$ with $\phi_{\mathbf{z}}$ denoting the Gaussian density function having the same first and second moments as $\mathbf{z}$. The quantity $\mathcal{J}(\mathbf{z})$, called the *negentropy* of $\mathbf{z}$, has the good property of being invariant with respect to invertible linear transformation, i.e. $\mathcal{J}(\mathbf{z}) = \mathcal{J}(\mathbf{W}\mathbf{x}) = \mathcal{J}(\mathbf{x})$ for any invertible matrix $\mathbf{W}$. On the other hand, if $\mathbf{z}$ satisfies $\mathrm{Cov}(\mathbf{z}) = \mathbf{I}$, as required in Definition 2.1.3, then the third term on the right hand side of (2.1.7) vanishes. We recall that for whitened signal $\mathbf{x}$, $\mathrm{Cov}(\mathbf{z}) = \mathbf{I}$ holds for any orthogonal matrix $\mathbf{W}$. Thus, if we take into account the preprocessing procedure, by (2.1.7) solving optimization problem (2.1.6) is equivalent to find an orthogonal matrix $\mathbf{W}^*$ that maximizes $\sum_{k=1}^{d} \mathcal{J}(z_i)$.

We note that $\mathcal{J}(z_k) = \mathcal{J}(\mathbf{w}_k^\mathsf{T}\mathbf{x})$ for $k = 1, \ldots, d$, where $\mathbf{w}_k$ is the k-th row of $\mathbf{W}$. One way to maximize the sum $\sum_{k=1}^{d} \mathcal{J}(\mathbf{w}_k^\mathsf{T}\mathbf{x})$, is to maximize each $\mathcal{J}(\mathbf{w}_k^\mathsf{T}\mathbf{x})$ separately. If we are able to find an orthonormal family of vectors $\mathbf{w}_1^*, \ldots, \mathbf{w}_d^*$ such that there holds in the local sense

$$\mathbf{w}_k^* = \operatorname*{argmax}_{\mathbf{w} \in \mathcal{S}} \mathcal{J}(\mathbf{w}^\mathsf{T}\mathbf{x}), \quad k = 1, \ldots, d, \tag{2.1.8}$$

then $\mathbf{W}^*$ formed by $\mathbf{w}_1^*, \ldots, \mathbf{w}_d^*$ as rows should be a demixing matrix. This intuition can be justified as follows. The negentropy $\mathcal{J}(z_k)$ is a measure of distance from z_k to Gaussianity by definition. Thus by maximizing $\mathcal{J}(\mathbf{w}^\mathsf{T}\mathbf{x})$ with respect to $\mathbf{w}$, we are actually searching a linear combination of $\mathbf{x}$ whose distribution is the most distant from Gaussian. In other words, we seek the "non-Gaussianity" (Hyvärinen & Oja, 2000) and we claim that "non-Gaussianity" leads to independent component. The primitive idea is that, by the central limit theorem, the average of several independent random variables tend to be Gaussian under certain conditions; hence, intuitively, if we can fabricate a random variable from linear combination of some independent components, such that it is the least Gaussian possible, then

the obtained random variable itself should coincide with one of the underlying independent components. This argument seems a bit too wild, but the conclusion can be made rigourous. The following result is a variation of Theorem 11 (page 58) in (Vrins, 2007):

Theorem 2.1.12. *Let* $\mathbf{s} = (s_1, \ldots, s_d)$ *be the source signal with independent components among which none is Gaussian. Then the mapping* $\mathbf{w} \in \mathcal{S} \to \mathcal{J}(\mathbf{w}^\mathsf{T}\mathbf{s})$ *reaches local maximum at* $\mathbf{w} = \pm\mathbf{e}_i$ *for* $i = 1, \ldots, d$.

Now let us examine (2.1.8). The maximization of $\mathcal{J}(\mathbf{w}^\mathsf{T}\mathbf{x})$ requires the calculation of an integral of type (2.1.5) where the probability density function of $\mathbf{x}$ is directly involved. This is not a easy task from a practical point of view. One way to overcome this difficulty is to use quantities that are more easily accessible such as moments or cumulants to approximate the negentropy. The following formula (Comon, 1994) is a classical approximation for z of zero mean and unit variance.:

$$\mathcal{J}(z) \approx \frac{1}{12}\kappa_3^2 + \frac{1}{48}\kappa_4^2 + \frac{7}{48}\kappa_3^4 - \frac{1}{8}\kappa_3^2\kappa_4, \tag{2.1.9}$$

where κ_i stands for the i-th order cumulant of the underlying the random variable. Approximations of type (2.1.9) have the drawback of being non robust against outliers in practice. An alternative approximation (Hyvärinen & Oja, 2000) is the following:

$$\mathcal{J}(z) \approx \sum_{i=1}^{n} c_i \Big(\mathbb{E}[G_i(z) - G_i(v)]\Big)^2, \tag{2.1.10}$$

where c_i are some positive constants, v is a standard Gaussian random variable, and the functions G_i are some nonquadratic functions. The advantage of (2.1.10) is that by choosing an appropriate G, we can obtain approximations of negentropy that are better than the one given in (2.1.9). Note that the term on the right hand side of (2.1.10) is not always a valid approximation of the negentropy, but the term by itself can always be used as a measure of non-Gaussianity that is consistent in the sense that it is always non negative, and equal to zero if z has a Gaussian distribution.

The simplest case of (2.1.10) is $n = 1$, where we have

$$\mathcal{J}(z) \propto \Big(\mathbb{E}[G(z) - G(v)]\Big)^2. \tag{2.1.11}$$

If we are to maximize $\mathcal{J}(z) = \mathcal{J}(\mathbf{w}^\mathsf{T}\mathbf{x})$ subject to $\|\mathbf{w}\| = 1$ using approximation 2.1.11, we need only to maximize or minimize $\mathbb{E}[G(\mathbf{w}^\mathsf{T}\mathbf{x})]$. In fact, for any $\mathbf{w}$ having unit norm, we always have $\mathbb{E}[\mathbf{w}^\mathsf{T}\mathbf{x}] = 0$ and $\mathrm{Cov}(\mathbf{w}^\mathsf{T}\mathbf{x}) = 1$ by Assumption 2. Therefore, the Gaussian random variable v which by definition has the same first and second moments as $z = \mathbf{w}^\mathsf{T}\mathbf{x}$, is independent of the choice of $\mathbf{w}$. We then deduce that $\mathbb{E}[G(v)]$ is a constant for fixed G.

Now that approximation (2.1.11) is a quadratic function of $\mathbb{E}[G(\mathbf{w}^\mathsf{T}\mathbf{x})]$, it reaches its maximum if and only if $\mathbb{E}[G(\mathbf{w}^\mathsf{T}\mathbf{x})]$ is maximized or minimized.

Due to reasons stated above, in this thesis we will only consider the following type of contrast function:

Definition 2.1.13. *Let $G(\cdot)$ be a twice continuously differentiable nonlinear and nonquadratic function, referred to as the* **nonlinearity**, *and $\mathbf{x}$ be the observed signal. The function $G(\cdot, \mu) : \mathcal{S} \to \mathbb{R}$ defined by*

$$G(\mathbf{w}, \mu) \stackrel{\text{def}}{=} \mathbb{E}_\mu[G(\mathbf{w}^\mathsf{T}\mathbf{x})], \quad \mathbf{w} \in \mathcal{S} \tag{2.1.12}$$

is called the **contrast function**.

Remark 2.1.14. In the notation of contrast function, we use the same letter G to indicate its connection with the nonlinearity function $G(\cdot)$. The second argument μ of $G(\mathbf{w}, \mu)$ refers to the underlying probability distribution with respect to which the mathematical expectation is taken. Whenever there is no risk of confusion, the theoretical probability distribution μ in the notation $\mathbb{E}_\mu[\cdot]$ is omitted for simplicity.

Remark 2.1.15. In order to be consistent with the notation used in (Hyvärinen & Oja, 2000; Hyvärinen, 1999), we write $g(x) \stackrel{\text{def}}{=} G'(x)$, the derivative of $G(x)$. By abuse of language, both $g(\cdot)$ and $G(\cdot)$ will be referred to as the "nonlinearity function". Besides, in order to distinguish from the empirical contrast function that will be defined in Chapter 4, we will sometimes call $G(\mathbf{w}, \mu)$ the *theoretical* contrast function.

The following theorem (Hyvärinen & Oja, 2000) confirms that the contrast function $G(\cdot, \mu)$ defined in (2.1.12) can be utilized as a criterion for ICA.

Theorem 2.1.16. *Consider model 2.1.1 with Assumption 1 and 2. and let $\mathbf{a}_i$ be the i-th column vector of the mixing matrix $\mathbf{A}$. Then $\pm\mathbf{a}_i$ is a local minimizer or maximizer of the contrast function $G(\cdot, \mu)$ on the unit sphere $\mathcal{S}$ for $i = 1, \ldots, d$, provided that*

$$\mathbb{E}[g'(s_i) - g(s_i)s_i] \neq 0. \tag{2.1.13}$$

Remark 2.1.17. The condition (2.1.13) is consistent with the requirement that the source signals $s_1, \ldots, s_d$ must not be Gaussian. In fact, if this is the case for s_i, then we have

$$\begin{aligned}
\mathbb{E}[g'(s_i) - g(s_i)s_i] &= \int_{\mathbb{R}} \phi(x)\Big(g'(x) - g(x)x\Big)\mathrm{d}x \\
&= \int_{\mathbb{R}} \phi(x)g'(x)\mathrm{d}x - \int_{\mathbb{R}} \phi(x)g(x)x\mathrm{d}x \\
&= \phi(x)g(x)\Big|_{-\infty}^{\infty} - \int_{\mathbb{R}} g(x)\mathrm{d}\phi(x) - \int_{\mathbb{R}} \phi(x)g(x)x\mathrm{d}x,
\end{aligned}$$

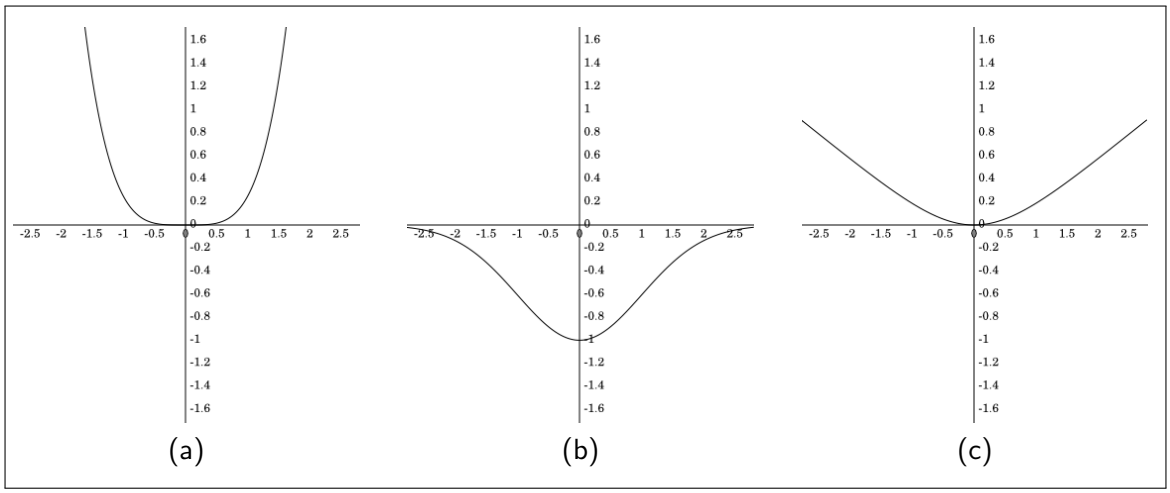

Figure 2.3: Graphs of the three popular nonlinearity functions. (a) "kurtosis", (b) "Gauss", (c) "tanh".

where to deduce the last equality we used integration by parts. Notice that for standard Gaussian density $\phi(x) = \exp(-x^2/2)/\sqrt{2\pi}$, there holds $\mathrm{d}\phi(x) = -x\phi(x)\mathrm{d}x$, hence the last two terms are cancelled out. Moreover, if $g(x)$ can be bounded by a polynomial function, then the first term vanishes as well. We then deduce that $\mathbb{E}[g'(s_i) - g(s_i)s_i] = 0$.

Remark 2.1.18. Theorem 2.1.16 tells us that all sources can be found by optimizing the contrast function (2.1.12), as long as we search the optimizer in the neighbourhood of a demixing vector; however, it does not guarantee that all local maximizers and minimizers correspond necessarily to a demixing vector.

The choice of nonlinearity functions can be quite flexible, and we only implicitly require that the mathematical expectation (2.1.12) is well-defined. Popular nonlinearity functions include the following:

"kurtosis": $G_1(x) = \frac{1}{4}x^4, \quad g_1(x) = x^3,$

"Gauss": $G_2(x) = -\exp(-\frac{x^2}{2}), \quad g_2(x) = x\exp(-\frac{x^2}{2})$

"tanh": $G_3(x) = \log\cosh(x), \quad g_3(x) = \tanh(x).$

Nonlinearity G_1 is referred to as "kurtosis" due to its obvious relation with the true kurtosis $\kappa_4(z) = \mathbb{E}[z^4] - 3$ (fourth order cumulant) of a standardized random variable z. Kurtosis based contrast function can date back as early as the invention of ICA (Donoho, 1981). The other two nonlinearities, G_2 and G_3, were first proposed in (Hyvärinen, 1999) along with the FastICA

algorithm. Contrast functions based on the latter two nonlinearities have the advantage of being more robust against outliers.

Example 2.1.19. We plotted the contrast functions in the 2-dimensional case based on three popular nonlinearities: "kurtosis", "Gauss" and "tanh". Two different scenarios are considered: (i), Two source signals have different distributions (one uniform and one Laplace); (ii), Two source distributions are the same (both uniform). In this simplest case, according to Theorem 2.1.16, there should be exactly 4 demixing vectors, namely $\pm\mathbf{a}_1, \pm\mathbf{a}_2$, which are either local maximizers or minimizers of the contrast function. Therefore, if the contrast function is turned out to have more than 4 local optima, then there must exist spurious solution of the demixing vector. From the figure, we observe that when the two source signals have different distributions, then the corresponding contrast functions possess exactly 4 optima (2 global maxima and 2 global minima). On the contrary, if the two source distributions are all uniform, then in all the three cases there exist 8 optima. This means that 4 optima do not correspond to a demixing vector.

2.1.4 FastICA algorithm

As we have seen in the last section, independent components can be recovered by optimizing the contrast function $G(\mathbf{w}, \mu)$ subject to the constraint $\|\mathbf{w}\| = 1$. In principle, we can either try to find an algebraic solution to the original optimization problem, or we can use an adaptive method to generate a sequence that converges to the true solution. However, in most cases, an analytic closed-form solution for ICA problem does not exist, hence we must resort to the adaptive method.

This thesis aims at giving a rigorous analysis of a popular adaptive method called *FastICA*, also known as *The fixed-point algorithm*. It is one of the most successful algorithms for independent component analysis in terms of accuracy and low computational complexity. It was first proposed by Hyvarinen and Oja from Finnish school (Hyvärinen, 1999) in late 90s. There are two versions of FastICA: *one-unit FastICA* and *symmetric FastICA*. As the name suggests, one-unit FastICA corresponds to the one-unit separation, which estimates one row at a time of the demixing matrix $\mathbf{W}$, while symmetric FastICA (Oja, 2002; Oja & Yuan, 2006) corresponds to the simultaneous separation which estimates $\mathbf{W}$ as a whole. The analysis of symmetric FastICA is beyond the scope of this thesis, and we will hereby concentrate on the one-unit version of FastICA.

In what follows, a nonlinearity function $g = G'$ is supposed to be fixed. The original form of one-unit FastICA algorithm can be stated as follows:

Algorithm 2.1.20 (One-unit FastICA for extraction of one source)**.**

1. *Choose an arbitrary initial point* $\mathbf{w} \in \mathcal{S}$.

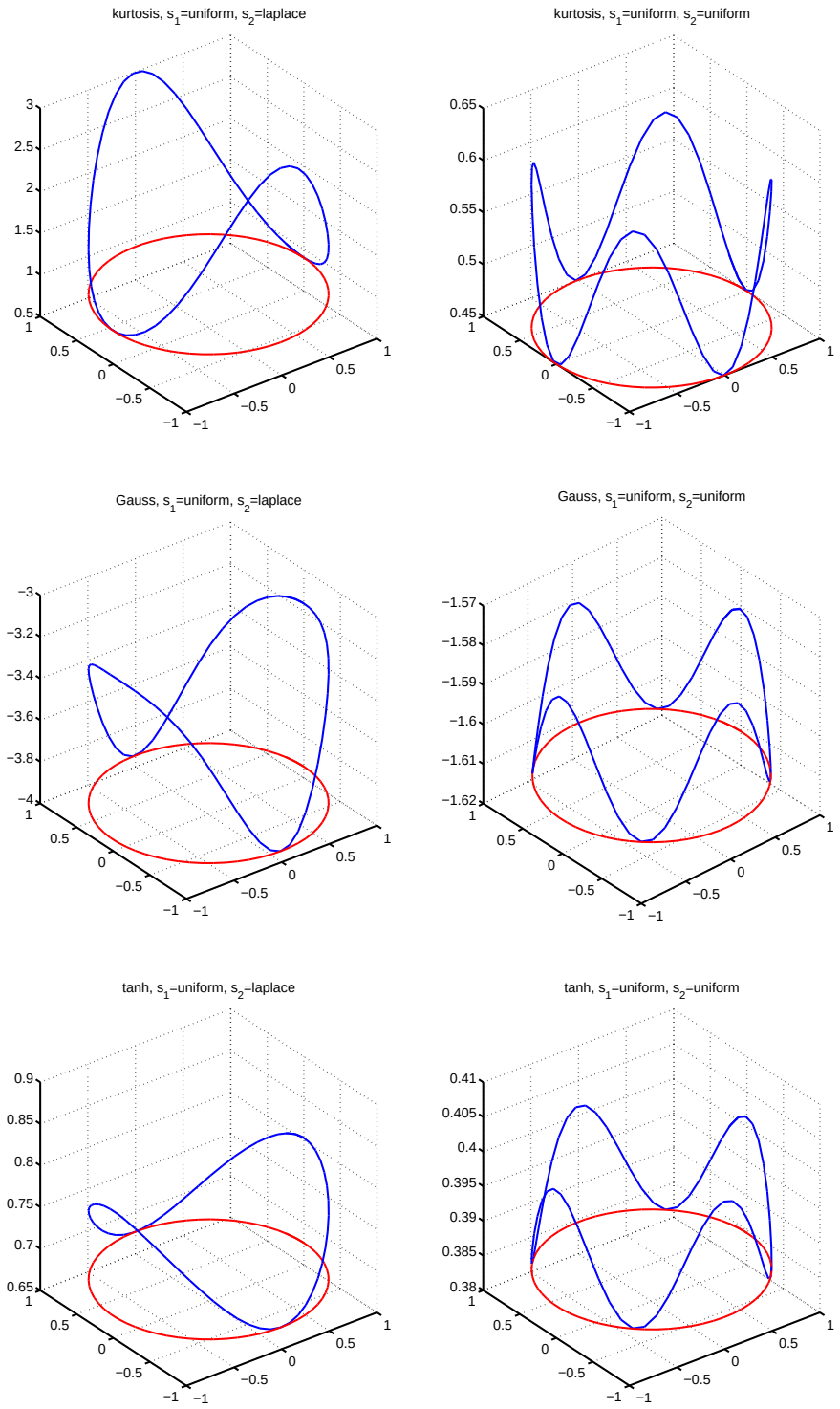

Figure 2.4: Contrast functions in the 2-dimensional case based on three popular nonlinearities. On the left, the underlying two source signals are respectively uniform and Laplace; On the right, two source distributions are both uniform.

2. Run the following iteration until convergence:

$$\mathbf{w}^+ \leftarrow \mathbb{E}[g'(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{w} - g(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{x}]$$
$$\mathbf{w} \leftarrow \frac{\mathbf{w}^+}{\|\mathbf{w}^+\|}.$$

FastICA algorithm 2.1.20 was derived initially as an approximate Newton method applied to the optimization problem

$$\min_{\mathbf{w}\in\mathcal{S}} \mathbb{E}[G(\mathbf{w}^\mathsf{T}\mathbf{x})] \quad \text{or} \quad \max_{\mathbf{w}\in\mathcal{S}} \mathbb{E}[G(\mathbf{w}^\mathsf{T}\mathbf{x})]. \tag{2.1.14}$$

By the method of Lagrange multipliers, we know that optima of $\mathbb{E}[G(\mathbf{w}^\mathsf{T}\mathbf{x})]$ subject to the constraint $\|\mathbf{w}\| = 1$ can be obtained by setting the first order derivative of the corresponding Lagrange function $\mathscr{L}(\mathbf{w}, \lambda)$ to zero, where

$$\mathscr{L}(\mathbf{w}, \lambda) \stackrel{\text{def}}{=} \mathbb{E}[G(\mathbf{w}^\mathsf{T}\mathbf{x})] + \frac{\lambda}{2}(\|\mathbf{w}\|^2 - 1).$$

Now we try to solve

$$\partial_{\mathbf{w}}\mathscr{L}(\mathbf{w}, \lambda) = \mathbb{E}[g(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{x}] + \lambda\mathbf{w} = 0 \tag{2.1.15}$$

using Newton's method, where $\partial_{\mathbf{w}}$ denotes the partial derivative with respect to $\mathbf{w}$. We recall that Newton's method is a iterative scheme that can be used to find numerically the roots of a smooth real valued function $F : \mathbb{R}^d \to \mathbb{R}^d$.

Algorithm 2.1.21 (Newton's method)**.**

1. Choose an initial estimate $\mathbf{y}_0 \in \mathbb{R}^d$.

2. Calculate

$$\mathbf{y}_{i+1} = \mathbf{y}_i - \big(F'(\mathbf{y}_i)\big)^{-1} F(\mathbf{y}_i) \tag{2.1.16}$$

for $i = 1, 2, \ldots$ *until convergence is achieved.*

Now let us define $F(\mathbf{w}) = \partial_{\mathbf{w}}\mathscr{L}(\mathbf{w}, \lambda)$ and take λ as a constant. In order to apply Algorithm 2.1.21, we need to invert the Jacobian matrix F', where by (2.1.15)

$$F'(\mathbf{w}) = \mathbb{E}[g'(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{x}\mathbf{x}^\mathsf{T}] + \lambda\mathbf{I}. \tag{2.1.17}$$

To simplify the inversion of this matrix, we use the approximation

$$\mathbb{E}[g'(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{x}\mathbf{x}^\mathsf{T}] \approx \mathbb{E}[g'(\mathbf{w}^\mathsf{T}\mathbf{x})]\mathbb{E}[\mathbf{x}\mathbf{x}^\mathsf{T}] = \mathbb{E}[g'(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{I}], \tag{2.1.18}$$

where to deduce the last equality we used the fact that the observed signal $\mathbf{x}$ is whitened. Note that if we use approximation (2.1.18), then the Jacobian matrix $F'(\mathbf{w})$ becomes diagonal:

$$F'(\mathbf{w}) \approx \mathbb{E}[g'(\mathbf{w}^\mathsf{T}\mathbf{x}) + \lambda]\mathbf{I},$$

hence it can be easily inverted. It then follows that

$$F'(\mathbf{w}) \approx \Big(\mathbb{E}[g'(\mathbf{w}^\mathsf{T}\mathbf{x}) + \lambda]\Big)^{-1}\mathbf{I}. \tag{2.1.19}$$

Using (2.1.15) and (2.1.19) to (2.1.16), we obtain the following approximative Newton iteration:

$$\mathbf{w}^+ = \mathbf{w} - \frac{\mathbb{E}[g(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{x}] + \lambda\mathbf{w}}{\mathbb{E}[g'(\mathbf{w}^\mathsf{T}\mathbf{x}) + \lambda]}. \tag{2.1.20}$$

Note that (2.1.20) can be further simplified:

$$\begin{aligned}\mathbf{w}^+ &= \frac{1}{\mathbb{E}[g'(\mathbf{w}^\mathsf{T}\mathbf{x}) + \lambda]}\Big(\mathbf{w}\mathbb{E}[g'(\mathbf{w}^\mathsf{T}\mathbf{x}) + \lambda] - \big(\mathbb{E}[g(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{x}] + \lambda\mathbf{w}\big)\Big) \\ &= \frac{1}{\mathbb{E}[g'(\mathbf{w}^\mathsf{T}\mathbf{x}) + \lambda]}\Big(\mathbb{E}[g'(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{w} - g(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{x}]\Big).\end{aligned} \tag{2.1.21}$$

Since we shall eventually project $\mathbf{w}^+$ back to the unit sphere $\mathcal{S}$ by multiplying $1/\|\mathbf{w}^+\|$, coefficient $1/\mathbb{E}[g'(\mathbf{w}^\mathsf{T}\mathbf{x}) + \lambda]$ in (2.1.21) can be removed. This gives the FastICA iteration:

$$\begin{aligned}\mathbf{w}^+ &= \mathbb{E}[g'(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{w} - g(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{x}] \\ \mathbf{w} &= \frac{\mathbf{w}^+}{\|\mathbf{w}^+\|}.\end{aligned}$$

Remark 2.1.22. The heuristic derivation given above shows how the FastICA algorithm is inspired from Newton's method, but it does not explain why the algorithm should work. Theoretical result (Hyvärinen, 1999) that guarantees the validity of FastICA is Theorem 2.1.23 stated below, and we will discuss it in detail in Chapter 3. Here, we just point out that the most important advantages of FastICA over the original Newton's method is that while retaining a quadratic convergence speed as Newton's method, the FastICA algorithm does not require the inversion of any matrix. This makes its computational complexity significantly lower than that of Newton's method. Besides, unlike Newton's method where bad initial estimate may result in the failure of convergence, numerical experiments indicate that FastICA performs generally very well regardless of the initial point $\mathbf{w}_0$.

Theorem 2.1.23 ((Hyvärinen, 1999)). *Consider model 2.1.1 with Assumption 1 and 2. Let $\mathbf{a}_i$ be the i-th column vector of the mixing matrix $\mathbf{A}$ such that*

$$\mathbb{E}[g'(s_i) - g(s_i)s_i] \neq 0. \tag{2.1.22}$$

Then there exists $r > 0$, such that if $\mathbf{w}_0 \in \mathcal{S} \cap \mathscr{B}_r(\mathbf{a}_i)$, the sequence generated by the FastICA algorithm 2.1.20 converges to $\mathbf{a}_i$.

Theorem 2.1.23 states that FastICA algorithm converges to some column of the mixing matrix (which is a demixing vector), but it is not known in advance which colomn the algorithm finds. The limit of FastICA mainly depends on the neighborhood $\mathscr{B}_r(\mathbf{a}_i)$ within which lies the initial iterate. If we wish to find more than one demixing vector without the algorithm converging to the same vector twice, a decorrelation or "deflation" constraint (Delfosse & Loubaton, 1995) must be added: at each step, the i-th estimated vector must be perpendicular to the $i-1$ previously found vectors, since the demixing matrix should be orthogonal. This is usually achieved by using a Gram-Schmidt type of orthogonalization method :

Algorithm 2.1.24 (One-unit FastICA for extraction of d sources)**.**

1. *Set* $p = 0$.
2. *Choose an arbitrary initial point* $\mathbf{w} \in \mathcal{S}$.
3. *Run the following iteration until convergence:*

$$\begin{aligned} \mathbf{w}^+ &= \mathbb{E}[g'(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{w} - g(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{x}] \\ \mathbf{w}^+ &= \mathbf{w}^+ - \sum_{i=1}^{p} \mathbf{w}_i \mathbf{w}_i^\mathsf{T} \mathbf{w}^+. \\ \mathbf{w} &= \frac{\mathbf{w}^+}{\|\mathbf{w}^+\|}. \end{aligned} \tag{2.1.23}$$

4. *Write* $\mathbf{w}_{p+1} = \mathbf{w}$. *If* $p < d-1$, *set* $p = p+1$ *then go back to step 2; stop the algorithm otherwise.*

Note that the difference between Algorithm 2.1.20 and Algorithm 2.1.24 is essentially procedure (2.1.23). Analysis of this additional deflation procedure is beyond the scope of our work, and we will concentrate on Algorithm 2.1.20. Note that Step 2 of Algorithm 2.1.20 can be represented by the iteration of the following mapping $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^d$:

$$\mathbf{w} \to \frac{\mathbb{E}_\mu[g'(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{w} - g(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{x}]}{\|\mathbb{E}_\mu[g'(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{w} - g(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{x}]\|}. \tag{2.1.24}$$

Unsurprisingly, many properties of the FastICA algorithm can be revealed by studying mapping (2.1.24). Thus, it deserves a name in its own right.

Definition 2.1.25. *For* $\mathbf{w} \in \mathbb{R}^d$, *we define*

$$\boldsymbol{h}(\mathbf{w}, \mu) \stackrel{\text{def}}{=} \mathbb{E}_\mu[g'(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{w} - g(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{x}], \tag{2.1.25}$$

$$\boldsymbol{f}(\mathbf{w}, \mu) \stackrel{\text{def}}{=} \frac{\boldsymbol{h}(\mathbf{w}, \mu)}{\|\boldsymbol{h}(\mathbf{w}, \mu)\|}. \tag{2.1.26}$$

The mapping $\boldsymbol{f}(\cdot, \mu) : \mathbb{R}^d \to \mathbb{R}^d$ *is called the* **FastICA function**.

Using this notation, we can rewrite algorithm 2.1.20 as follows:

Algorithm 2.1.26.

1. *Choose an arbitrary initial point* $\mathbf{w}_0$ *on the unit sphere* $\mathcal{S}$.
2. *Run the following iteration until convergence:*

$$\mathbf{w} \leftarrow \boldsymbol{f}(\mathbf{w}, \mu).$$

FastICA function (2.1.26) will sometimes be called the **theoretical FastICA function**. The term "theoretical" is added to highlight its underlying theoretical ICA model, where we have perfect knowledge of the distribution μ and hence all the mathematical expectations involved can be precisely evaluated. Likewise, FastICA Algorithm 2.1.20 or 2.1.26 may be referred to as the **theoretical FastICA algorithm**. However, when there is no risk of confusion, the term "theoretical" is sometimes omitted for brevity.

Aside from results given in Theorem 2.1.23, the theoretical FastICA algorithm is proven to possess the following properties:

- It has locally at least a quadratic convergence speed (Hyvärinen, 1999). In the following particular cases, the convergence speed is even cubic (Shen, Kleinsteuber, & Hüper, 2008):
  - The nonlinearity is "kurtosis";
  - The nonlinearity $G(\cdot)$ is a even function and the extracted source signal s_i has a symmetrical distribution;
  - All the sources other than the extracted source s_i have a symmetrical distribution.
- The convergence is monotonic (Regalia & Kofidis, 2003). More precisely, for a FastICA generated sequence $\{\mathbf{w}_n\}$ that converges to $\mathbf{a}_i$ for some i, the sequence $\{G(\mathbf{w}_n, \mu)\}$ converges monotonically to $G(\mathbf{a}_i, \mu)$.

2.2 Empirical ICA

2.2.1 Empirical ICA Model

In practice, people work with a finite number of independent and identically distributed (i.i.d.) realizations of the observed signal $\mathbf{x}$. More precisely, only a finite sequence $\mathbf{x}(1), \ldots, \mathbf{x}(N)$ is available with each $\mathbf{x}(t)$ issued from model 2.1.1:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \quad t = 1, \ldots, N, \tag{2.2.1}$$

where

- The index t is the realization label. All the realizations are independent and identically distributed .
- The source signals $\mathbf{s}(1), \ldots, \mathbf{s}(N)$ are unknown, non-Gaussian and d-dimensional. Moreover, they have independent components with unit variance.
- The observed signals $\mathbf{x}(1), \ldots, \mathbf{x}(N)$ are known, while their probability distribution μ is not.
- The unknown mixing matrix $\mathbf{A}$ is a non-singular square matrix.

Remark 2.2.1. Assumption 1 on page 9 is already taken into account in the description above, while Assumption 2 on page 11 is not. That is, we do suppose that the source signal is not Gaussian and has unit variance, but for now it can have non-zero mean and the mixing matrix need not be orthogonal.

In what follows, (2.2.1) will be referred to as the empirical ICA model. The aim of empirical ICA becomes estimating the mixing matrix $\mathbf{W}$.

Problem 2.2.2 (Empirical ICA problem)**.** *The empirical ICA problem consists of giving an estimation* $\widehat{\mathbf{W}}$ *of the demixing matrix* $\mathbf{W}$ *based upon the observation* $\mathbf{x}(1), \ldots, \mathbf{x}(N)$.

2.2.2 Probability measure based on observation data

Now let us consider the empirical ICA problem 2.2.2. In order to implement the FastICA algorithm, one must do the following:

- Make sure that the observed signal $\mathbf{x}$ have zero mean and unit variance;
- Find way to evaluate the FastICA function (2.1.26).

As explained in Section 2.1.2, the first task can be done by centering and whitening the observed signal $\mathbf{x}(1), \ldots, \mathbf{x}(N)$. Although the exact value of $\mathbb{E}[\mathbf{x}]$ and $\mathrm{Cov}(\mathbf{x})$ needed to carry out data preprocessing are not known, these quantities can always be estimated by the sample mean and the sample variance respectively :

$$\mathbb{E}[x] \approx \bar{\mathbf{x}} \overset{\text{def}}{=} \frac{1}{N} \sum_{t=1}^{N} \mathbf{x}(t) \tag{2.2.2}$$

$$\mathrm{Cov}(\mathbf{x}) \approx \mathbf{C}_N \overset{\text{def}}{=} \frac{1}{N} \sum_{t=1}^{N} \mathbf{x}(t)\mathbf{x}(t)^{\mathsf{T}} - \bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathsf{T}}. \tag{2.2.3}$$

Using (2.2.2) and (2.2.3), we can represent the centered and whitened data as (similar to 2.1.3)

$$\begin{aligned}\tilde{\mathbf{x}}(t) &= \left(\frac{1}{N}\sum_{t=1}^{N}\mathbf{x}(t)\mathbf{x}(t)^T - \bar{\mathbf{x}}\bar{\mathbf{x}}^T\right)^{-\frac{1}{2}}\left(\mathbf{x}(t) - \frac{1}{N}\sum_{t=1}^{N}\mathbf{x}(t)\right)\\ &= \mathbf{C}_N^{-\frac{1}{2}}\Big(\mathbf{x}(t) - \bar{\mathbf{x}}\Big), \quad t = 1, \ldots, N.\end{aligned}$$

Clearly, the preprocessed data $\tilde{\mathbf{x}}(1), \ldots, \tilde{\mathbf{x}}(N)$ has the following properties:

- zero sample mean: $\dfrac{1}{N}\displaystyle\sum_{t=1}^{N}\tilde{\mathbf{x}}(t) = 0.$
- unit sample variance:

$$\frac{1}{N}\sum_{t=1}^{N}\Big(\tilde{\mathbf{x}}(t) - \frac{1}{N}\sum_{t=1}^{N}\tilde{\mathbf{x}}(t)\Big)\Big(\tilde{\mathbf{x}}(t) - \frac{1}{N}\sum_{t=1}^{N}\tilde{\mathbf{x}}(t)\Big)^{\mathsf{T}} = \mathbf{I}.$$

We then assert that the preprocessed data $\tilde{\mathbf{x}}(1), \ldots, \tilde{\mathbf{x}}(N)$ can be used to implement the FastICA algorithm.

As for the second task, we can follow the same route. Although the FastICA function (2.1.25) cannot be directly evaluated due to the fact that we do not know μ, it can always be estimated by an appropriate estimator. As what we did in (2.2.2) and (2.2.3), the sample average is a natural candidate. We first calculate

$$\boldsymbol{h}(\mathbf{w}, \mu) \approx \frac{1}{N}\sum_{t=1}^{N}\Big(g'(\mathbf{w}^{\mathsf{T}}\tilde{\mathbf{x}}(t))\mathbf{w} - g(\mathbf{w}^{\mathsf{T}}\tilde{\mathbf{x}}(t))\tilde{\mathbf{x}}(t)\Big),$$

and then project it to $\mathcal{S}$ to obtain an estimate of $\boldsymbol{f}(\mathbf{w}, \mu)$.

These ideas lead us to consider the following discrete measures (i.e. distributions) constructed upon the observation $\mathbf{x}(1), \ldots, \mathbf{x}(N)$:

$$\mu_N^1 = \frac{1}{N}\sum_{t=1}^{N}\delta_{\mathbf{x}(t)} \tag{2.2.4}$$

$$\mu_N^2 = \frac{1}{N}\sum_{t=1}^{N}\delta_{(\mathbf{x}(t)-\bar{\mathbf{x}})} \tag{2.2.5}$$

$$\mu_N^3 = \frac{1}{N}\sum_{t=1}^{N}\delta_{\mathbf{Q}_N^{-1/2}\mathbf{x}(t)} \tag{2.2.6}$$

$$\mu_N^4 = \frac{1}{N}\sum_{t=1}^{N}\delta_{\mathbf{C}_N^{-1/2}(\mathbf{x}(t)-\bar{\mathbf{x}})} \tag{2.2.7}$$

where

$$\mathbf{Q}_N = \frac{1}{N}\sum_{t=1}^{N}\mathbf{x}(t)\mathbf{x}(t)^T.$$

Clearly, distributions (2.2.4)-(2.2.7) satisfy the property $\mu_N^k(\mathbb{R}) = 1$ for $k = 1, 2, 3, 4$, hence they are all probability distributions. Then for any function $f(\cdot)$, we can define the mathematical expectation of $f(\mathbf{z})$ with respect to the distribution μ_N^k of $\mathbf{z}$:

$$\mathbb{E}_{\mu_N^k}[f(\mathbf{z})] \stackrel{\text{def}}{=} \int f(\mathbf{z})\mu_N^k(\mathrm{d}\mathbf{z}) = \frac{1}{N}\sum_{t=1}^{N} f(\mathbf{z}(t)), \tag{2.2.8}$$

where $\mathbf{z}(1), \ldots, \mathbf{z}(N)$ denotes the support of $\mathbf{z}$. Thanks to (2.2.8), the mathematical expectation operator $\mathbb{E}_{\mu_N^k}[\cdot]$ can be used to denote the average of $f(\cdot)$ evaluated at $\mathbf{z}(1), \ldots, \mathbf{z}(N)$. Note that for each $k = 1, 2, 3, 4$, definition (2.2.8) can be explicitly written as:

$$\begin{aligned}
\mathbb{E}_{\mu_N^1}[f(\mathbf{z})] &\stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{w}^{\mathsf{T}}\mathbf{x}(t)) \\
\mathbb{E}_{\mu_N^2}[f(\mathbf{z})] &\stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{w}^{\mathsf{T}}(\mathbf{x}(t) - \bar{\mathbf{x}})) \\
\mathbb{E}_{\mu_N^3}[f(\mathbf{z})] &\stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{w}^{\mathsf{T}}\mathbf{Q}_N^{-1/2}\mathbf{x}(t)) \\
\mathbb{E}_{\mu_N^4}[f(\mathbf{z})] &\stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{w}^{\mathsf{T}}\mathbf{C}_N^{-1/2}(\mathbf{x}(t) - \bar{\mathbf{x}})).
\end{aligned}$$

It is easy to see that distributions (2.2.4)-(2.2.7) have specific meanings. More precisely, μ_N^1 stands for the classical empirical measure arising from the sample $\mathbf{x}(1), \ldots, \mathbf{x}(N)$, while μ_N^2, μ_N^3 and μ_N^4 can respectively be considered as the "empirical measure" based upon the centered data $\{\mathbf{x}(k) - \bar{\mathbf{x}}\}$, the whitened data $\{\mathbf{Q}_N^{-1/2}\mathbf{x}(k)\}$ and the centered and whitened data $\{\mathbf{C}_N^{-1/2}(\mathbf{x}(k) - \bar{\mathbf{x}})\}$. Of course, the utility of these distributions depend on the particular assumption of the model. In the most general case, where the source signal may have non-zero mean and the mixing matrix is arbitrary (with full rank), the data preprocessing is necessary and hence only μ_N^4 is meaningful. Nevertheless, on occasion people encounter, for example, signals that have intrinsically zero mean. In this case, the centering procedure can be omitted, then both μ_N^3 and μ_N^4 are valid for ICA. Likewise, if the observed signal is naturally uncorrelated, then the whitening procedure is no longer needed and we may therefore consider μ_N^2 and μ_N^4. As a summary, we distinguish the following situations:

1. $\mathbb{E}[\mathbf{x}] = 0$ and $\mathrm{Cov}(\mathbf{x}) = \mathbf{I}$: $\mu_N^1 - \mu_N^4$ are all suitable.
2. $\mathbb{E}[\mathbf{x}] \neq 0$ and $\mathrm{Cov}(\mathbf{x}) = \mathbf{I}$: only μ_N^2 and μ_N^4 are suitable.
3. $\mathbb{E}[\mathbf{x}] = 0$ and $\mathrm{Cov}(\mathbf{x}) \neq \mathbf{I}$: only μ_N^3 and μ_N^4 are suitable.
4. $\mathbb{E}[\mathbf{x}] \neq 0$ and $\mathrm{Cov}(\mathbf{x}) \neq \mathbf{I}$: only μ_N^4 is suitable.

In Chapter 4 and 5 that are dedicated to empirical ICA, we consider only the first situation and study all four measures $\mu_N^1 - \mu_N^4$. We claim that, although it seems to be the most particular case at first glance, we do not actually lose any generality. As a matter of fact, constructing μ_N^4 requires data centering and whitening regardless the actual mean and variance of the observed signal. Therefore, when studying μ_N^4, we always work with centered and whitened signal, as such their original mean and variance are irrelevant. For this reason, we do not exploit the properties $\mathbb{E}[\mathbf{x}] = 0$ and $\mathrm{Cov}(\mathbf{x}) = \mathbf{I}$ during our study, and the latter assumptions are merely made for the possibility of comparing the "performance" of all four measures.

2.2.3 Empirical FastICA algorithm

Using distribution μ_N^k, we are now able to define the empirical FastICA function, and hence the empirical FastICA algorithm. The empirical FastICA function is essentially a generalization of the theoretical one, because we only replace the measure μ in Definition 2.1.25 by μ_N^k for $k = 1, 2, 3, 4$:

Definition 2.2.3. *For $\mathbf{w} \in \mathbb{R}^d$ and $k = 1, 2, 3, 4$, we define*

$$\boldsymbol{h}(\mathbf{w}, \mu_N^k) \stackrel{\text{def}}{=} \mathbb{E}_{\mu_N^k}\Big[g'(\mathbf{w}^\mathsf{T}\mathbf{z})\mathbf{w} - g(\mathbf{w}^\mathsf{T}\mathbf{z})\mathbf{z}\Big] \tag{2.2.9}$$

$$\boldsymbol{f}(\mathbf{w}, \mu_N^k) \stackrel{\text{def}}{=} \frac{\boldsymbol{h}(\mathbf{w}, \mu_N^k)}{\|\boldsymbol{h}(\mathbf{w}, \mu_N^k)\|}. \tag{2.2.10}$$

We call $\boldsymbol{f}(\cdot, \mu_N^k)$ the **empirical FastICA function** *with respect to μ_N^k, or simply the empirical FastICA function.*

Similar to the theoretical case, the empirical FastICA algorithm is a scheme of self-iteration of $\boldsymbol{f}(\cdot, \mu_N^k)$:

Algorithm 2.2.4 (Empirical FastICA)**.**

1. *Choose an arbitrary initial point $\mathbf{w}$ on the unit sphere $\mathcal{S}$.*
2. *Run the following iteration until convergence:*

$$\mathbf{w} \leftarrow \boldsymbol{f}(\mathbf{w}, \mu_N^k).$$

The definition of Algorithm 2.2.4 does not guarantee its convergence. Although numerical simulation suggests that the convergence does hold, and it is *a priori* admitted by many authors, a rigorous prove of its convergence is still mixing in the community. The eager to fill this blank is the starting point of the whole work, and this task will be accomplished in Chapter 4.

Chapter 3

Theoretical FastICA Algorithm

Contents

This chapter is intended to reestablish the classical results concerning the theoretical FastICA algorithm. We will study the link between the critical points of the contrast function and the convergence of the FastICA algorithm. We will show that the columns of the mixing matrix are local minimizers of the contrast function and prove the relation $\mathbf{a} \in \mathrm{Min}\big(G(\mathbf{w},\mu)\big) \subset \mathrm{Fix}\big(\boldsymbol{f}(\mathbf{w},\mu)\big)$, where $\mathrm{Min}\big(G(\mathbf{w},\mu)\big)$ and $\mathrm{Fix}\big(\boldsymbol{f}(\mathbf{w},\mu)\big)$ denotes respectively the set of local minimizers of the contrast function $G(\mathbf{w},\mu)$ on the unit sphere and the set of fixed points of the FastICA function $\boldsymbol{f}(\mathbf{w},\mu)$. Moreover, we will show that the FastICA algorithm converges with at least a quadratic convergence speed to each column $\mathbf{a}$ of the mixing matrix.

3.1 Assumptions and method

Throughout this Chapter, we consider the theoretical ICA model (2.1.1) with Assumption 1 and 2, i.e.

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{3.1.1}$$

where

- The observed signal $\mathbf{x}$ has probability distribution μ with $\mathbb{E}[\mathbf{x}] = 0$ and $\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathsf{T}}] = \mathbf{I}$.

- The source signal $\mathbf{s}$ has independent, non-Gaussian components with $\mathbb{E}[\mathbf{s}] = 0$ and $\mathbb{E}[\mathbf{s}\mathbf{s}^\mathsf{T}] = \mathbf{I}$.
- The mixing matrix $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_d)$ is orthogonal, i.e. $\mathbf{A}\mathbf{A}^\mathsf{T} = \mathbf{I}$.

Aside from the basic assumptions listed above, in our analysis we need the following additional hypotheses:

Assumption 3. *(i) The nonlinearity function G has continuous derivatives up to the fourth order. Moreover, there exits $p > 0$ such that $G^{(k)}(t) \leq c|t|^p$ for $k = 0, \ldots, 4$, where c is some positive constant.*

(ii) The random vector $\mathbf{x}$ has finite moment of any order.

(iii) The function $H(\cdot, \mu) : \mathbb{R}^d \to \mathbb{R}$ defined by

$$H(\mathbf{w}, \mu) \stackrel{\text{def}}{=} \mathbb{E}_\mu[g'(\mathbf{w}^\mathsf{T}\mathbf{x}) - g(\mathbf{w}^\mathsf{T}\mathbf{x})(\mathbf{w}^\mathsf{T}\mathbf{x})] \tag{3.1.2}$$

satisfies $H(\mathbf{a}, \mu) > 0$.

We claim that none of the assumptions above is restrictive. First, it is easily seen that three most popular nonlinearity functions, namely "kurtosis", i.e. $g(x) = x^3$, "Gauss", i.e. $g(x) = -x\exp(-x^2/2)$ and "tanh", i.e. $g(x) = \tanh(x)$ satisfy assumption (i). As for assumption (ii), we claim that it was made in its current form to lighten the proof, and can be easily weakened. In fact, we require only that $\mathbf{x}$ has finite moment up to some l-th order, with l depending on p. Lastly, we point out that convergence of the FastICA algorithm relies on the necessary condition $H(\mathbf{a}, \mu) \neq 0$, see (Hyvärinen, 1999). To develop a rigorous convergence analysis of the FastICA algorithm, one needs to avoid the well-known sign-flipping phenomenon, i.e. FastICA oscillates between neighborhoods of two antipodes on the unit sphere, which causes the discontinuity of the corresponding FastICA map on the unit sphere. Although one can overcome this difficulty by generalizing the notion of algorithm convergence, or by using the concept of principal fiber bundles (Shen et al., 2008), we choose to make the convention that $H(\mathbf{a}, \mu) > 0$ to ensure that the convergence takes place in the traditional sense. Assumption (iii) has the advantage of being simple and always feasible for one-unit FastICA, and one needs only to choose the appropriate sign for the underlying nonlinearity function. The following remark reveals the connection of (3.1.2) with Hermit polynomial, which leads us to adopt the notation $H(\cdot, \cdot)$.

Remark 3.1.1. Let us define $\mathcal{H}_n(\cdot, \cdot) : \mathcal{C}^n \times \mathbb{R} \to \mathbb{R}$:

$$\mathcal{H}_n(g, t) = \frac{1}{\gamma(t)} \frac{\mathrm{d}^n(\gamma\, g)(t)}{\mathrm{d}^n t},$$

where $\gamma(t) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2})$. We remark that $\mathcal{H}_{n+1}(g, t) = \mathcal{H}_n(\mathcal{H}_1(g, \cdot), t)$, and $\mathcal{H}_n(1, t) = (-1)^n \mathcal{H}_n(t)$ where $\mathcal{H}_n$ is the Hermit polynomial of degree n.

In particular, the Hermit polynomial of degree 4 is $x^4 - 6x^2 + 3$, which is essentially the kurtosis when regarded as the nonlinearity function. Starting from a non-linearity function g, we get the sequence of non-linearity functions $\mathcal{H}_n(g,t)$, $n = 1, \ldots$. Finally, the function $H(\mathbf{w}, \mu)$ defined in (3.1.2) can be written as $H(\mathbf{w}, \mu) = \mathbb{E}_\mu[\mathcal{H}_1(g, \mathbf{w}^\mathsf{T}\mathbf{x})]$. More generally, let us denote $H_n(g, \mathbf{w}, \mu) = \mathbb{E}_\mu[\mathcal{H}_n(g, \mathbf{w}^\mathsf{T}\mathbf{x})]$. We will see at least numerically that a is a local minimizer of $H_{2n}(g, \mathbf{w}, \mu)$, but it is a local maximizer of $H_{2n+1}(g, \mathbf{w}, \mu)$.

In this chapiter, the following orthogonal projection method is frequently used. For any $\mathbf{w} \in \mathcal{S}$, we denote by $\Pi_\mathbf{w}$ the matrix of the orthogonal projection from $\mathbb{R}^d$ to span$(\mathbf{w})$ and by $\Pi_\mathbf{w}^\perp$ the matrix of the orthogonal projection from $\mathbb{R}^d$ to span$(\mathbf{w})^\perp$. Clearly, we have

$$\Pi_\mathbf{w} = \mathbf{w}\mathbf{w}^\mathsf{T}, \tag{3.1.3}$$

$$\Pi_\mathbf{w}^\perp = \mathbf{I} - \mathbf{w}\mathbf{w}^\mathsf{T}. \tag{3.1.4}$$

For any $\mathbf{x} \in \mathbb{R}^d$, we have the orthogonal decomposition

$$\mathbf{x} = \Pi_\mathbf{w}\mathbf{x} + \Pi_\mathbf{w}^\perp\mathbf{x}, \tag{3.1.5}$$

The following result shows that the decomposition (3.1.5) is vital in our analysis.

Lemma 3.1.2. *Let* $\mathbf{x}$ *be the signal defined in (3.1.1) and* $\mathbf{a}$ *be a column of the mixing matrix* $\mathbf{A}$*. Then* $\Pi_\mathbf{a}\mathbf{x}$ *and* $\Pi_\mathbf{a}^\perp\mathbf{x}$ *are independent random vectors.*

Proof. Let us write $\mathbf{A} = (\mathbf{a}_1, \cdots \mathbf{a}_d)$ and consider its ith column vector $\mathbf{a}_i$. First, we show that $\mathbf{a}_i^\mathsf{T}\mathbf{x}$ and $\mathbf{x} - \mathbf{a}_i(\mathbf{a}_i^\mathsf{T}\mathbf{x})$ are independent. Since $\mathbf{A}$ is orthogonal, we have $\mathbf{a}_i^\mathsf{T}\mathbf{A} = \mathbf{e}_i^\mathsf{T}$. It follows that $\mathbf{a}_i^\mathsf{T}\mathbf{x} = \mathbf{a}_i^\mathsf{T}\mathbf{A}\mathbf{s} = s_i$. On the other hand, we have also $\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_{j=1}^d \mathbf{a}_j s_j$. Hence

$$\mathbf{x} - \mathbf{a}_i(\mathbf{a}_i^\mathsf{T}\mathbf{x}) = \sum_{j=1}^d \mathbf{a}_j s_j - \mathbf{a}_i s_i = \sum_{i \neq j} \mathbf{a}_i s_i.$$

By the hypothesis that $\mathbf{s}$ has independent components, we get the independence between $\mathbf{a}^\mathsf{T}\mathbf{x}$ and $(\mathbf{I} - \mathbf{a}\mathbf{a}^\mathsf{T})\mathbf{x}$. □

Remark 3.1.3. Lemma 3.1.2 is a direct consequence of the fundamental hypothesis of ICA. It states that the observed signal $\mathbf{x}$ can be decomposed into a sum of two independent and perpendicular signals. Inversely, for a random vector $\mathbf{x}$, if there exists an orthonormal set of d vectors $\mathbf{a}_1, \ldots, \mathbf{a}_d$ such that $\mathbf{a}_i\mathbf{a}_i^\mathsf{T}\mathbf{x}$ and $(\mathbf{I} - \mathbf{a}_i\mathbf{a}_i^\mathsf{T})\mathbf{x}$ are independent, then there exists a matrix $\mathbf{A} \overset{\text{def}}{=} (\mathbf{a}_1, \ldots, \mathbf{a}_d)$ and a random vector $\mathbf{s}$ with independent components such that $\mathbf{x} = \mathbf{A}\mathbf{s}$. Hence, Lemma 3.1.2 completely characterizes the ICA model 2.1.1

3.2 Minimizers of contrast function and fixed points of FastICA

Proposition 3.2.1. *For any* $\mathbf{w}, \mathbf{v} \in \mathcal{S}$*, we have*

$$G(\mathbf{w},\mu) = G(\mathbf{v},\mu) + (\mathbf{w}-\mathbf{v})^{\mathsf{T}}\boldsymbol{\varphi}(\mathbf{v},\mu) + \frac{1}{2}(\mathbf{w}-\mathbf{v})^{\mathsf{T}}\mathbf{K}(\mathbf{v},\mu)(\mathbf{w}-\mathbf{v}) + \mathcal{O}(\|\mathbf{w}-\mathbf{v}\|^3),$$

where $\boldsymbol{\varphi}(\mathbf{v},\mu)$ *and* $\mathbf{K}(\mathbf{v},\mu)$ *are defined by*

$$\boldsymbol{\varphi}(\mathbf{v},\mu) \stackrel{\text{def}}{=} \mathbb{E}[g(\mathbf{v}^{\mathsf{T}}\mathbf{x})\Pi_{\mathbf{v}}^{\perp}\mathbf{x}], \tag{3.2.1}$$

$$\mathbf{K}(\mathbf{v},\mu) \stackrel{\text{def}}{=} H(\mathbf{v},\mu)\mathbf{I} + \mathbb{E}[g'(\mathbf{v}^{\mathsf{T}}\mathbf{x})\Pi_{\mathbf{v}}^{\perp}(\mathbf{x}\mathbf{x}^{\mathsf{T}} - \mathbf{I})\Pi_{\mathbf{v}}^{\perp}]. \tag{3.2.2}$$

Proof. See Section 3.5.1.

Lemma 3.2.2. *A vector* $\mathbf{v}$ *is a fixed point of the FastICA function if and only if* $\boldsymbol{\varphi}(\mathbf{v},\mu) = 0$ *and* $H(\mathbf{v},\mu) > 0$*.*

Proof. By definition, vector $\mathbf{v}$ is a fixed point of $\boldsymbol{f}(\cdot,\mu)$ if and only if $\boldsymbol{f}(\mathbf{v},\mu) = \mathbf{v}$. Note that

$$\begin{aligned}
\boldsymbol{f}(\mathbf{v},\mu) &= \frac{1}{\|\boldsymbol{h}(\mathbf{v},\mu)\|}\mathbb{E}[g'(\mathbf{v}^{\mathsf{T}}\mathbf{x})\mathbf{v} - g(\mathbf{v}^{\mathsf{T}}\mathbf{x})\mathbf{x}] \\
&= \frac{1}{\|\boldsymbol{h}(\mathbf{v},\mu)\|}\mathbb{E}[g'(\mathbf{v}^{\mathsf{T}}\mathbf{x})\mathbf{v} - g(\mathbf{v}^{\mathsf{T}}\mathbf{x})\Pi_{\mathbf{v}}\mathbf{x} - g(\mathbf{v}^{\mathsf{T}}\mathbf{x})\Pi_{\mathbf{v}}^{\perp}\mathbf{x}] \\
&= \frac{1}{\|\boldsymbol{h}(\mathbf{v},\mu)\|}\Big(H(\mathbf{v},\mu)\mathbf{v} - \boldsymbol{\varphi}(\mathbf{v},\mu)\Big),
\end{aligned}$$

where the term $\boldsymbol{\varphi}(\mathbf{v},\mu) = \mathbb{E}[g(\mathbf{v}^{\mathsf{T}}\mathbf{x})\Pi_{\mathbf{v}}^{\perp}\mathbf{x}]$ is perpendicular to $\mathbf{v}$. Therefore $\boldsymbol{f}(\mathbf{v},\mu)$ is parallel to $\mathbf{v}$ if and only if $\boldsymbol{\varphi}(\mathbf{v},\mu) = 0$. Note that in this case we have

$$\|\boldsymbol{h}(\mathbf{v},\mu)\| = \Big(\|H(\mathbf{v},\mu)\mathbf{v}\|^2 + \|\boldsymbol{\varphi}(\mathbf{v},\mu)\|^2\Big)^{1/2} = |H(\mathbf{v},\mu)|.$$

Then $\boldsymbol{f}(\mathbf{v},\mu) = \mathbf{v}$ implies $H(\mathbf{v},\mu) > 0$. □

Remark 3.2.3. We clarify that a point $\mathbf{v}$ being a fixed point of the FastICA function $\boldsymbol{f}(\mathbf{v},\mu)$ means it satisfies $\mathbf{v} = \boldsymbol{f}(\mathbf{v},\mu)$, and it does not need to be the limit of the FastICA *algorithm*. In the next section, we will see that other condition is needed for the FastICA algorithm to converge to $\mathbf{v}$. In this thesis, we will avoid using the statement like "fixed point of the FastICA algorithm" since it makes confusion.

From Proposition 3.2.1 and Lemma 3.2.2, we deduce immediately the following result.

Proposition 3.2.4. *If* $\mathbf{v}$ *is a fixed point of the FastICA function and if the matrix* $\mathbf{K}(\mathbf{v},\mu)$ *is positive definite, then* $\mathbf{v}$ *is a local minimizer of the contrast function.*

Proposition 3.2.5. *If* $\mathbf{v}$ *is a local minimizer of the contrast function* $G(\mathbf{w},\mu)$ *on* $\mathcal{S}$*, and if* $H(\mathbf{v},\mu)>0$*, then it is a fixed point of the FastICA function.*

Proof. From Taylor's formula, we have

$$G(\mathbf{w},\mu)=G(\mathbf{v},\mu)+(\mathbf{w}-\mathbf{v})^T\mathbb{E}[g(\mathbf{v}^T\mathbf{x})\mathbf{x}]+\mathcal{O}(\|\mathbf{w}-\mathbf{v}\|^2),$$

or equivalently

$$\frac{G(\mathbf{w},\mu)-G(\mathbf{v},\mu)}{\|\mathbf{w}-\mathbf{v}\|}=\frac{(\mathbf{w}-\mathbf{v})^T\mathbb{E}[g(\mathbf{v}^T\mathbf{x})\mathbf{x}]}{\|\mathbf{w}-\mathbf{v}\|}+\mathcal{O}(\|\mathbf{w}-\mathbf{v}\|). \quad (3.2.3)$$

On the one hand, we can show that

$$\left\{\mathbf{u}:\mathbf{u}=\lim_{\mathbf{w}\in\mathcal{S}\to\mathbf{v}}\frac{\mathbf{w}-\mathbf{v}}{\|\mathbf{w}-\mathbf{v}\|}\right\}=\operatorname{span}(\mathbf{v}^\perp).$$

On the other hand, if $\mathbf{v}$ is a local minimizer of the contrast function on $\mathcal{S}$, then for all $\mathbf{w}\in\mathcal{S}$ near $\mathbf{v}$ we have

$$\frac{G(\mathbf{w},\mu)-G(\mathbf{v},\mu)}{\|\mathbf{w}-\mathbf{v}\|}\geq 0.$$

Applying this result to (3.2.3) and letting $\mathbf{w}\to\mathbf{v}$, we obtain

$$\lim_{\mathbf{w}\in\mathcal{S}\to\mathbf{v}}\mathbb{E}\Big[\frac{(\mathbf{w}-\mathbf{v})^Tg(\mathbf{v}^\mathsf{T}\mathbf{x})\mathbf{x}}{\|\mathbf{w}-\mathbf{v}\|}\Big]\geq 0.$$

Hence, we have $\mathbb{E}[g(\mathbf{v}^T\mathbf{x})\mathbf{u}^T\mathbf{x}]\geq 0$ for all $\mathbf{u}\in\operatorname{span}(\mathbf{v}^\perp)$, which implies that $\boldsymbol{\varphi}(\mathbf{v},\mu)=\mathbb{E}[g(\mathbf{v}^T\mathbf{x})\boldsymbol{\Pi}_{\mathbf{v}^\perp}\mathbf{x}]=0$. This condition along with the hypothesis $H(\mathbf{v},\mu)>0$ gives $\mathbf{v}=\boldsymbol{f}(\mathbf{v},\mu)$. □

Proposition 3.2.6. *The vector* $\mathbf{a}$ *is a fixed point of the FastICA function. It is also a local minimizer of the contrast function on* $\mathcal{S}$*.*

Proof. Let us show that $\mathbf{a}$ being a column of $\mathbf{A}$ implies

$$\boldsymbol{\varphi}(\mathbf{a},\mu)=0 \quad \text{and} \quad \mathbf{K}(\mathbf{a},\mu)=H(\mathbf{a},\mu)\mathbf{I}.$$

By Lemma 3.1.2, the random vectors $\mathbf{a}^\mathsf{T}\mathbf{x}$ and $(\mathbf{I}-\mathbf{a}\mathbf{a}^\mathsf{T})\mathbf{x}$ are independent. Then it follows from the assumption $\mathbb{E}[\mathbf{x}]=0$ that

$$\boldsymbol{\varphi}(\mathbf{a},\mu)=\mathbb{E}[g(\mathbf{a}^\mathsf{T}\mathbf{x})(\mathbf{I}-\mathbf{a}\mathbf{a}^\mathsf{T})\mathbf{x}]=\mathbb{E}[g(\mathbf{a}^\mathsf{T}\mathbf{x})]\mathbb{E}[(\mathbf{I}-\mathbf{a}\mathbf{a}^\mathsf{T})\mathbf{x}]=0.$$

To prove the second, let us denote $\mathbf{L}(\mathbf{w},\mu)\overset{\text{def}}{=}\mathbb{E}[g'(\mathbf{w}^\mathsf{T}\mathbf{x})\Pi_{\mathbf{v}}^{\perp}(\mathbf{x}\mathbf{x}^\mathsf{T}-\mathbf{I})\Pi_{\mathbf{v}}^{\perp}]$. Using the decomposition $\mathbf{x}=\mathbf{a}\mathbf{a}^\mathsf{T}\mathbf{x}+(\mathbf{I}-\mathbf{a}\mathbf{a}^\mathsf{T})\mathbf{x}$, we get

$$\begin{aligned}
\mathbf{L}(\mathbf{a},\mu)&=\mathbb{E}[g'(\mathbf{a}^\mathsf{T}\mathbf{x})(\mathbf{x}\mathbf{x}^\mathsf{T}-\mathbf{a}\mathbf{a}^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\mathbf{a}\mathbf{a}^\mathsf{T}-\mathbf{I}+\mathbf{a}\mathbf{a}^\mathsf{T})]\\
&=\mathbb{E}\Big[g'(\mathbf{a}^\mathsf{T}\mathbf{x})\Big((\mathbf{a}\mathbf{a}^\mathsf{T}\mathbf{x}+(\mathbf{I}-\mathbf{a}\mathbf{a}^\mathsf{T})\mathbf{x})(\mathbf{a}\mathbf{a}^\mathsf{T}\mathbf{x}+(\mathbf{I}-\mathbf{a}\mathbf{a}^\mathsf{T})\mathbf{x})^\mathsf{T}\\
&\quad-\mathbf{a}\mathbf{a}^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\mathbf{a}\mathbf{a}^\mathsf{T}-\mathbf{I}+\mathbf{a}\mathbf{a}^\mathsf{T}\Big)\Big]\\
&=\mathbb{E}\Big[g'(\mathbf{a}^\mathsf{T}\mathbf{x})\Big(\mathbf{a}\mathbf{a}^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}(\mathbf{I}-\mathbf{a}\mathbf{a}^\mathsf{T})+(\mathbf{I}-\mathbf{a}\mathbf{a}^\mathsf{T})\mathbf{x}\mathbf{x}\mathbf{a}\mathbf{a}^\mathsf{T}\\
&\quad+(\mathbf{I}-\mathbf{a}\mathbf{a}^\mathsf{T})\mathbf{x}\mathbf{x}^\mathsf{T}(\mathbf{I}-\mathbf{a}\mathbf{a}^\mathsf{T})-\mathbf{I}+\mathbf{a}\mathbf{a}^\mathsf{T}\Big)\Big]. \qquad (3.2.4)
\end{aligned}$$

Note that we have

$$\begin{aligned}&\mathbb{E}\big[g'(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a}\mathbf{a}^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}(\mathbf{I}-\mathbf{a}\mathbf{a}^\mathsf{T})\big]\\ =\ &\mathbb{E}\big[g'(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a}\mathbf{a}^\mathsf{T}\mathbf{x}\big]\,\mathbb{E}\big[\mathbf{x}^\mathsf{T}(\mathbf{I}-\mathbf{a}\mathbf{a}^\mathsf{T})\big]=0\end{aligned}\tag{3.2.5}$$

by the independence between $\mathbf{a}^\mathsf{T}\mathbf{x}$ and $(\mathbf{I}-\mathbf{a}\mathbf{a}^\mathsf{T})\mathbf{x}$, and

$$\mathbb{E}\big[(\mathbf{I}-\mathbf{a}\mathbf{a}^\mathsf{T})\mathbf{x}\mathbf{x}^\mathsf{T}(\mathbf{I}-\mathbf{a}\mathbf{a}^\mathsf{T})\big]=\mathbf{I}-\mathbf{a}\mathbf{a}^\mathsf{T}\tag{3.2.6}$$

by the assumption $\mathbb{E}[\mathbf{x}\mathbf{x}^\mathsf{T}]=\mathbf{I}$. Applying (3.2.5) and (3.2.6) to (3.2.4), we obtain $\mathbf{L}(\mathbf{a},\mu)=0$ and $\mathbf{K}(\mathbf{a},\mu)=H(\mathbf{a},\mu)\mathbf{I}$.

Finally, we deduce from Lemma 3.2.2 that $\mathbf{a}$ is a fixed point of the FastICA function, and from Proposition 3.2.4 that it is also a local minimizer of the contrast function. □

3.3 Local Convergence of the FastICA Algorithm

Proposition 3.3.1. *Let* $\mathbf{v}$ *be a fixed point of the FastICA function* $\boldsymbol{f}(\cdot,\mu)$. *If* $\|\nabla\boldsymbol{f}(\mathbf{v},\mu)\|<1$, *then starting near* $\mathbf{v}$, *the FastICA algorithm converges to* $\mathbf{v}$.

Proof. Since $\|\nabla\boldsymbol{f}(\mathbf{v},\mu)\|<1$, by the continuity of $\nabla\boldsymbol{f}(\cdot,\mu)$, there exists $0<K<1$ and $r>0$, such that $\sup_{\mathbf{w}\in\mathscr{B}_r(\mathbf{v})}\|\nabla\boldsymbol{f}(\mathbf{w})\|<K$. Hence, for $\mathbf{w}_0\in\mathscr{B}_r(\mathbf{v})$, we have

$$\|\mathbf{w}_1-\mathbf{v}\|=\|\boldsymbol{f}(\mathbf{w}_0,\mu)-\boldsymbol{f}(\mathbf{v},\mu)\|\le K\cdot\|\mathbf{w}_0-\mathbf{v}\|.\tag{3.3.1}$$

It follows that $\|\mathbf{w}_n-\mathbf{v}\|\le K^n\cdot\|\mathbf{w}_0-\mathbf{v}\|$. Consequently, $\{\mathbf{w}_n\}$ is a Cauchy sequence that converges to $\mathbf{v}$. □

Lemma 3.3.2. *For all* $\mathbf{w}\in\mathcal{S}$, *we have*

$$\nabla\boldsymbol{h}(\mathbf{w},\mu)=\mathbb{E}_\mu[g''(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{w}\mathbf{x}^\mathsf{T}+g'(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{I}-g'(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{x}\mathbf{x}^\mathsf{T}],$$

$$\nabla\boldsymbol{f}(\mathbf{w},\mu)=\frac{\big(\|\boldsymbol{h}(\mathbf{w},\mu)\|^2\mathbf{I}-\boldsymbol{h}(\mathbf{w},\mu)\boldsymbol{h}(\mathbf{w},\mu)^\mathsf{T}\big)\nabla\boldsymbol{h}(\mathbf{w},\mu)}{\|\boldsymbol{h}(\mathbf{w},\mu)\|^3}.$$

Proposition 3.3.3. *Let* $\mathbf{a}$ *be a column of the mixing matrix* $\mathbf{A}$ *such that* $H(\mathbf{a},\mu)$ *defined in (3.1.2) is not zero. Then we have*

$$\boldsymbol{h}(\mathbf{a},\mu)=H(\mathbf{a},\mu)\mathbf{a}\tag{3.3.2}$$

$$\nabla\boldsymbol{h}(\mathbf{a},\mu)=\gamma(\mathbf{a},\mu)\mathbf{a}\mathbf{a}^\mathsf{T}\tag{3.3.3}$$

$$\boldsymbol{f}(\mathbf{a},\mu)=\mathbf{a}\tag{3.3.4}$$

$$\nabla\boldsymbol{f}(\mathbf{a},\mu)=0,\tag{3.3.5}$$

where $\gamma(\cdot,\mu)$ *is some scalar valued function.*

Proof.

(i). we have

$$\begin{aligned}\boldsymbol{h}(\mathbf{a},\mu) &= \mathbb{E}_\mu[g'(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a} - g(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{x}] \\ &= \mathbb{E}_\mu[g'(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a} - g(\mathbf{a}^\mathsf{T}\mathbf{x})\Pi_\mathbf{a}\mathbf{x}] - \mathbb{E}_\mu[g(\mathbf{a}^\mathsf{T}\mathbf{x})\Pi_\mathbf{a}^\perp\mathbf{x}].\end{aligned}$$

Since $\mathbf{a}$ is such that $\boldsymbol{\varphi}(\mathbf{a},\mu) = \mathbb{E}_\mu[g(\mathbf{a}^\mathsf{T}\mathbf{x})\Pi_\mathbf{a}^\perp\mathbf{x}] = 0$, we get

$$\begin{aligned}\boldsymbol{h}(\mathbf{a},\mu) &= \mathbb{E}_\mu[g'(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a} - g(\mathbf{a}^\mathsf{T}\mathbf{x})\Pi_\mathbf{a}\mathbf{x}] \\ &= \mathbb{E}_\mu[g'(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a} - g(\mathbf{a}^\mathsf{T}\mathbf{x})(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a}] \\ &= H(\mathbf{a},\mu)\mathbf{a},\end{aligned}$$

with $H(\mathbf{a},\mu) = \mathbb{E}_\mu[g'(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a} - \mathbf{a}^\mathsf{T}\mathbf{x}\, g(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a}]$.

(ii). Substituting $\mathbf{x}$ by $\Pi_\mathbf{a}\mathbf{x} + \Pi_\mathbf{a}^\perp\mathbf{x}$ in

$$\nabla\boldsymbol{h}(\mathbf{a},\mu) = \mathbb{E}_\mu[g''(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a}\mathbf{x}^\mathsf{T} + g'(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{I} - g'(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{x}\mathbf{x}^\mathsf{T}],$$

we obtain

$$\nabla\boldsymbol{h}(\mathbf{a},\mu) = \mathbb{E}_\mu[g''(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a}(\mathbf{x}^\mathsf{T}\Pi_\mathbf{a} + \mathbf{x}^\mathsf{T}\Pi_\mathbf{a}^\perp) + g'(\mathbf{a}^\mathsf{T}\mathbf{x})(\mathbf{a}\mathbf{a}^\mathsf{T} - \mathbf{a}\mathbf{a}^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\mathbf{a}\mathbf{a}^\mathsf{T})].$$

By the assumption on $\mathbf{a}$ and the fact that $\mathbb{E}_\mu[\mathbf{x}] = 0$, we have

$$\mathbb{E}_\mu[g''(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{x}^\mathsf{T}\Pi_\mathbf{a}^\perp] = \mathbb{E}_\mu[g''(\mathbf{a}^\mathsf{T}\mathbf{x})]\mathbb{E}_\mu[\mathbf{x}^\mathsf{T}\Pi_\mathbf{a}^\perp] = 0.$$

It follows that

$$\begin{aligned}\nabla\boldsymbol{h}(\mathbf{a},\mu) &= \mathbb{E}_\mu[g''(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a}(\mathbf{x}^\mathsf{T}\Pi_\mathbf{a}) + g'(\mathbf{a}^\mathsf{T}\mathbf{x})(\mathbf{a}\mathbf{a}^\mathsf{T} - \mathbf{a}\mathbf{a}^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\mathbf{a}\mathbf{a}^\mathsf{T})] \\ &= \mathbb{E}_\mu[g''(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a}\mathbf{x}^\mathsf{T}\mathbf{a}\mathbf{a}^\mathsf{T} + g'(\mathbf{a}^\mathsf{T}\mathbf{x})(\mathbf{a}\mathbf{a}^\mathsf{T} - \mathbf{a}\mathbf{a}^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\mathbf{a}\mathbf{a}^\mathsf{T})] \\ &= \mathbb{E}_\mu[g''(\mathbf{a}^\mathsf{T}\mathbf{x})(\mathbf{x}^\mathsf{T}\mathbf{a})\mathbf{a}\mathbf{a}^\mathsf{T} + g'(\mathbf{a}^\mathsf{T}\mathbf{x})(\mathbf{a}\mathbf{a}^\mathsf{T} - (\mathbf{x}^\mathsf{T}\mathbf{a})^2\mathbf{a}\mathbf{a}^\mathsf{T})] \\ &= \gamma(\mathbf{a},\mu)\mathbf{a}\mathbf{a}^\mathsf{T},\end{aligned}$$

where

$$\gamma(\mathbf{a},\mu) = \mathbb{E}_\mu[g''(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{x}^\mathsf{T}\mathbf{a} + g'(\mathbf{a}^\mathsf{T}\mathbf{x})(1 - (\mathbf{a}^\mathsf{T}\mathbf{x})^2)]. \qquad (3.3.6)$$

(iii). The equality $\boldsymbol{f}(\mathbf{a},\mu) = \mathbf{a}$ is a direct consequence of (i).

(iv). Then applying (i) and (ii) to

$$\nabla\boldsymbol{f}(\mathbf{w},\mu) = \frac{(\|\boldsymbol{h}(\mathbf{w},\mu)\|^2\mathbf{I} - \boldsymbol{h}(\mathbf{w},\mu)\boldsymbol{h}(\mathbf{w},\mu)^\mathsf{T})\nabla\boldsymbol{h}(\mathbf{w},\mu)}{\|\boldsymbol{h}(\mathbf{w},\mu)\|^3}$$

yields gives $\nabla\boldsymbol{f}(\mathbf{a},\mu) = 0$. □

Equalities (3.3.2) and (3.3.4) are just a consequence of the independence of the random vectors $\Pi_{\mathbf{a}}\mathbf{x}$ and $\Pi_{\mathbf{a}}^{\perp}\mathbf{x}$, while (3.3.3) and (3.3.5) are, to our knowledge, new. The fact $\boldsymbol{f}(\mathbf{a},\mu)=\mathbf{a}$ shows that the columns of the mixing matrix $\mathbf{A}$ are fixed points of the FastICA algorithm with respect to μ, while $\nabla\boldsymbol{f}(\mathbf{a},\mu)=0$ confirms that starting near $\mathbf{v}$, the FastICA algorithm converges to $\mathbf{v}$, according to Proposition 3.3.1. Moreover, $\nabla\boldsymbol{f}(\mathbf{a},\mu)=0$ implies that the sequence $\{\mathbf{w}_n\}$ generated by FastICA converges to $\mathbf{a}$ with a quadratic convergence speed. In fact, using Taylor's formula and taking into account of (3.3.4) (3.3.5), we obtain

$$\begin{aligned}\mathbf{w}_{n+1}&=\boldsymbol{f}(\mathbf{w}_n,\mu)=\boldsymbol{f}(\mathbf{a},\mu)+\nabla\boldsymbol{f}(\mathbf{a},\mu)(\mathbf{w}_n-\mathbf{a})+\mathcal{O}(\|\mathbf{w}_n-\mathbf{a}\|^2)\\&=\mathbf{a}+\mathcal{O}(\|\mathbf{w}_n-\mathbf{a}\|^2),\end{aligned}$$

where the lth entry of the vector $\mathcal{O}(\|\mathbf{w}_n-\mathbf{a}\|^2)$ equals to

$$\frac{1}{2}\sum_{1\le i,j\le d}\partial_{w_i}\partial_{w_j}\boldsymbol{f}_l(\boldsymbol{\xi}_l,\mu)(\mathbf{w}_n-\mathbf{a})_i(\mathbf{w}_n-\mathbf{a})_j.$$

which is bounded by

$$\frac{d}{2}\sup_{1\le i,j,l\le d,\mathbf{w}\in\mathscr{B}_r(\mathbf{a})}|\partial_{w_i}\partial_{w_j}\boldsymbol{f}_l(\mathbf{w},\mu)|\|\mathbf{w}_n-\mathbf{a}\|^2.$$

It follows that

$$\sup_n\frac{\|\mathbf{w}_{n+1}-\mathbf{a}\|}{\|\mathbf{w}_n-\mathbf{a}\|^2}\le+\infty. \tag{3.3.7}$$

The following theorem summarizes the discussion above.

Theorem 3.3.4. *There exists $r>0$, such that if $\mathbf{w}_0\in\mathcal{S}\cap\mathscr{B}_r(\mathbf{a})$, then the sequence generated by the FastICA algorithm converges to $\mathbf{a}$, i.e.*

$$\lim_{k\to\infty}\mathbf{w}_k=\mathbf{a},$$

Moreover, the convergence speed is quadratic.

Hyvarinen has already showed the quadratic speed of convergence of the FastICA algorithm (Hyvärinen & Oja, 1997; Hyvärinen, 1999). He proved that $\boldsymbol{h}(\mathbf{w})=\boldsymbol{h}(\mathbf{a})+\mathcal{O}(\|\mathbf{w}-\mathbf{a}\|^2)$ for all $\mathbf{w}\in\mathcal{S}\cap\mathscr{B}_r(\mathbf{a})$ for small r, and then he derived that $\boldsymbol{f}(\mathbf{w})=\boldsymbol{f}(\mathbf{a})+\mathcal{O}(\|\mathbf{w}-\mathbf{a}\|^2)$ for all $\mathbf{w}\in\mathcal{S}\cap\mathscr{B}_r(\mathbf{a})$. We emphasize that $\nabla\boldsymbol{f}(\mathbf{a})=0$ is not a direct consequence of the latter equality. For example we can show that

$$G(\mathbf{w},\mu)=G(\mathbf{a},\mu)+\mathcal{O}(|\mathbf{w}-\mathbf{a}|^2)$$

for $w\in\mathcal{S}\cap\mathscr{B}_r(\mathbf{a})$, but $\nabla G(a,\mu)\neq 0$. In our proof we show that $\nabla\boldsymbol{f}(\mathbf{a})=0$ and then $\boldsymbol{f}(\mathbf{w})=\boldsymbol{f}(\mathbf{a})+\mathcal{O}(\|\mathbf{w}-\mathbf{a}\|^2)$ for all $\mathbf{w}\in\mathscr{B}_r(\mathbf{a})$. Hence, our result shows that the term

$$\mathcal{O}(\|\mathbf{w}-\mathbf{a}\|^2)=\frac{1}{2}(\mathbf{w}-\mathbf{a})^T\nabla^2\boldsymbol{f}(\mathbf{v})(\mathbf{w}-\mathbf{a})$$

where $\mathbf{v}\in[\mathbf{a},\mathbf{w}]$.

Proposition 3.3.5. *If $G(x) = x^4$, then we have $\nabla^2 \boldsymbol{f}(\mathbf{a}, \mu) = 0$. As a result, the convergence speed of FastICA algorithm is cubic.*

Proof. See Section 3.5.2.

Hyvarinen and Oja have already showed the cubic speed of convergence of the FastICA algorithm for the kurtosis non-linearity (Hyvärinen, 1999; Oja, 2002). They proved that $h(\mathbf{w}) = h(\mathbf{a}) + \mathcal{O}(\|\mathbf{w} - \mathbf{a}\|^3)$ for all $\mathbf{w} \in \mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}$ for small r, and then he derived that $\boldsymbol{f}(\mathbf{w}) = \boldsymbol{f}(\mathbf{a}) + \mathcal{O}(\|\mathbf{w} - \mathbf{a}\|^3)$ for all $\mathbf{w} \in \mathcal{S} \cap \mathscr{B}_r(\mathbf{a})$. As we showed in the first commentary $\nabla \boldsymbol{f}(\mathbf{a}) = 0, \nabla^2 \boldsymbol{f}(\mathbf{a}) = 0$ is not a direct consequence of the latter equality. In our proof we show that $\nabla \boldsymbol{f}(\mathbf{a}) = 0, \nabla^2 \boldsymbol{f}(\mathbf{a}) = 0$ and then $\boldsymbol{f}(\mathbf{w}) = \boldsymbol{f}(\mathbf{a}) + \mathcal{O}(\|\mathbf{w} - \mathbf{a}\|^3)$ for all $\mathbf{w} \in \mathscr{B}_r(\mathbf{a})$. Hence, our result shows that the term

$$\mathcal{O}(\|\mathbf{w} - \mathbf{a}\|^3) = \frac{1}{6} \sum_{i,j,k} \partial^3_{w_i w_j w_k} \boldsymbol{f}(\mathbf{v})(\mathbf{w}_i - \mathbf{a}_i)(\mathbf{w}_j - \mathbf{a}_j)(\mathbf{w}_k - \mathbf{a}_k)$$

where $\mathbf{v}$ lies between $\mathbf{a}$ and $\mathbf{w}$.

3.4 Numerical results

3.4.1 Examples of contrast function and FastICA

In the sequel, we consider the case $d = 2$ that the two source signals s_1, s_2 have respectively Laplace and uniform distribution. We take kurtosis as the nonlinearity function, i.e. $G(x) = x^4$. Besides, without loss of generality, we suppose that $\mathbf{A} = \mathbf{I}$. It is clear that any vector $\mathbf{w} \in \mathcal{S}$ can be parameterized by a scalar $\theta \in [0, 2\pi)$ via $\mathbf{w}(\theta) = (\cos(\theta), \sin(\theta))^\mathsf{T}$, and hence the contrast function can be represented as a mapping

$$\theta \to G(\mathbf{w}(\theta), \mu) = \mathbb{E}[G(\cos(\theta)s_1 + \sin(\theta)s_2)]. \tag{3.4.1}$$

This approach, called *angular parametrization* (Vrins, 2007) is convenient for visualizing the numerical results in a 2D-plan. Note that we have $\mathbf{A} = (\mathbf{e}_1, \mathbf{e}_2)$ and $\mathbf{w}(0) = \mathbf{e}_1$, $\mathbf{w}(\pi/2) = \mathbf{e}_2$, $\mathbf{w}(\pi) = -\mathbf{e}_1$, $\mathbf{w}(3\pi/2) = -\mathbf{e}_2$.

Example 3.4.1. In Fig 3.1, we plot $G(\mathbf{w}(\theta, \mu))$ and $H(\mathbf{w}(\theta), \mu)$. We observe from the figure that, the contrast function attains its minimum at $\theta = \pi/2, 3\pi/2$, which correspond to $\pm\mathbf{e}_2$, and the function $H(\mathbf{w}, \mu)$ has positive value at both points. Inversely, at $\theta = 0, \pi$, or $\pm\mathbf{e}_1$, we have $H(\mathbf{w}, \mu) < 0$ and $G(\mathbf{w}, \mu)$ attains its local maximum. This example confirms Proposition 3.2.6.

Example 3.4.2. In Fig 3.2, we illustrate how FastICA algorithm converges to the local minimizer (or maximizer) of the contrast function. We iterated FastICA three times for input $\mathbf{w}_0(\theta) = (\cos(\theta), \sin(\theta))$ with θ ranging from 0 to π. We recorded the outcome of each iteration, namely

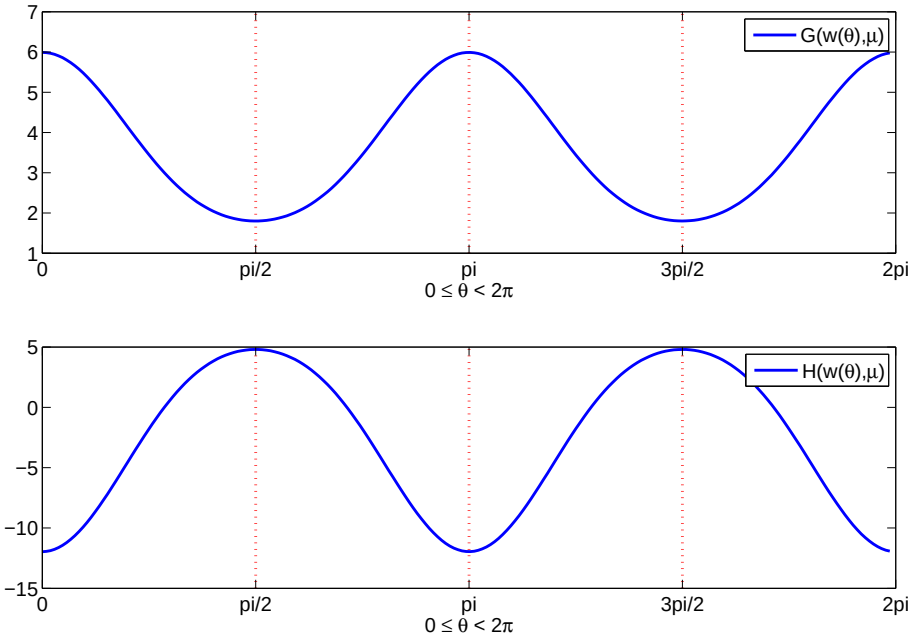

Figure 3.1: $G(\mathbf{w}(\theta), \mu)$ and $H(\mathbf{w}(\theta), \mu)$.

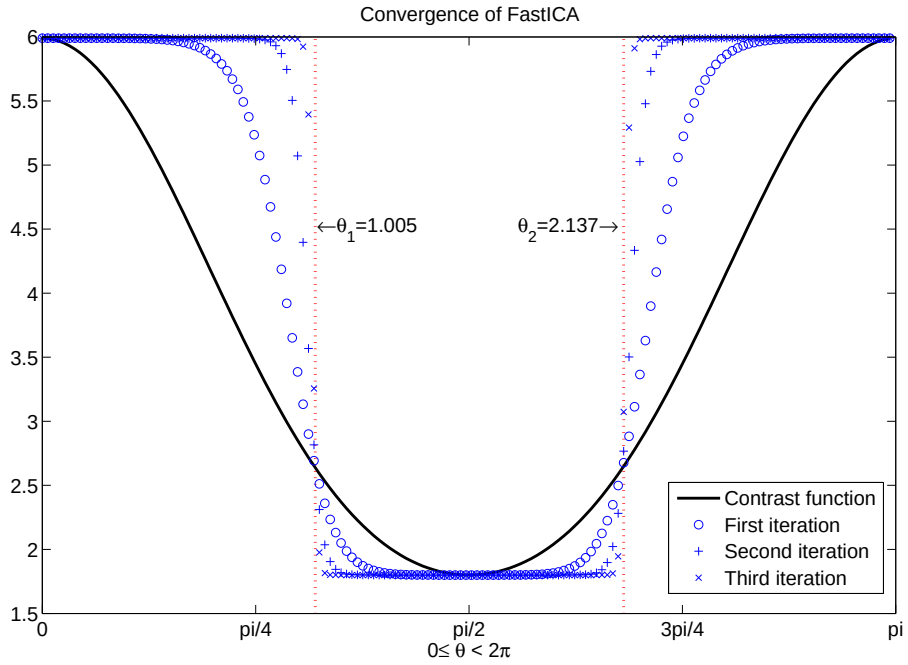

Figure 3.2: Convergence of FastICA.

$\mathbf{w}_i(\theta) \overset{\text{def}}{=} \boldsymbol{f}(\mathbf{w}_{i-1}(\theta), \mu)$ for $i = 1, 2, 3$ and for $0 \leq \theta \leq \pi$, and then plotted $G(\mathbf{w}_i(\theta), \mu)$, i.e. the contrast function evaluated at these points. In the figure, the black solid curve represents the contrast function (4.3.1), or equivalently $\theta \to G(\mathbf{w}_0(\theta), \mu)$ while the mark "$\circ$", "$+$" and "$\times$" stands respectively for the mapping $\theta \to G(\mathbf{w}_i(\theta), \mu)$ with $i = 1, 2, 3$. From the graph, we observe that for any initial input $\mathbf{w}_0(\theta)$, as the index i augments, $G(\mathbf{w}_i(\theta), \mu)$ tends to either $G(\mathbf{e}_1, \mu)$ or $G(\mathbf{e}_2, \mu)$ *monotonically*, the latter being a local minimum of $G(\mathbf{w}, \mu)$. A study of the monotonic convergence of the FastICA algorithm can be found in (Regalia & Kofidis, 2003). Moreover, if the angle θ of $\mathbf{w}_0$ lies within the interval $(1.005, 2.137)$, then $G(\mathbf{w}_i(\theta), \mu) \to G(\mathbf{e}_2, \mu)$.

Example 3.4.3. It is of interest to see among $\mathcal{S}$ which are fixed points of the FastICA function $\boldsymbol{f}(\mathbf{w}, \mu)$. By Lemma 3.2.2, fixed points are the solutions of the equation $\boldsymbol{\varphi}(\mathbf{w}, \mu) = 0$. In Fig 3.3, we plot the mapping $\theta \to \|\boldsymbol{\varphi}(\mathbf{w}(\theta), \mu)\|$ and $\theta \to \|\nabla \boldsymbol{f}(\mathbf{w}(\theta), \mu)\|$ for kurtosis nonlinearity function. We observe that $\|\boldsymbol{\varphi}(\mathbf{w}, \mu)\|$ has exactly 4 zeros which are $\pm\mathbf{e}_1$ and $\pm\mathbf{e}_2$. Therefore, any vector other than $\pm\mathbf{e}_1$ and $\pm\mathbf{e}_2$ cannot be fixed point of $\boldsymbol{f}(\mathbf{w}, \mu)$. Besides, we find out that $\|\nabla \boldsymbol{f}(\mathbf{w}(\theta), \mu)\|$ vanishes only at $\pm\mathbf{e}_1$ and $\pm\mathbf{e}_2$. In Fig 3.4, we plot the same mapping using Hermite polynomial $\mathcal{H}_6 = x^6 - 15x^4 + 45x^2 - 15$ as the nonlinearity function. From the graph, we observe that in this case, $\|\boldsymbol{\varphi}(\mathbf{w}(\theta), \mu)\|$ has 8 zeros including $\pm\mathbf{e}_1$ and $\pm\mathbf{e}_2$. $\|\nabla \boldsymbol{f}(\mathbf{w}(\theta), \mu)\|$ however, as in the case of kurtosis nonlinearity, it vanishes only at $\pm\mathbf{e}_1$ and $\pm\mathbf{e}_2$.

Example 3.4.4. Theoretically, the FastICA algorithm has a quadratic convergence speed for a general nonlinearity function, and in the case of kurtosis, as indicated in Corollary 3.3.5, the convergence speed is even cubic. It is of interest to see if this is really the case in the numerical simulations. In this example, we choose an arbitrary initial point $\mathbf{w}_0(\theta)$ such that the angle θ is near $\pi/2$. Starting from $\mathbf{w}_0(\theta)$, the FastICA algorithm yields a sequence $\{\mathbf{w}_n\}$ that converges to $\mathbf{w}(\pi/2) = \mathbf{e}_2$. In Fig 3.5 and 3.6, we plot respectively $\|\mathbf{w}_{n+1} - \mathbf{e}_2\|/\|\mathbf{w}_n - \mathbf{e}_2\|^2$ and $\|\mathbf{w}_{n+1} - \mathbf{e}_2\|/\|\mathbf{w}_n - \mathbf{e}_2\|$ for different n. From the graph, we see that even for the kurtosis nonlinearity, the ratio $\|\mathbf{w}_{n+1} - \mathbf{e}_2\|/\|\mathbf{w}_n - \mathbf{e}_2\|^2$ explodes immediately while $\|\mathbf{w}_{n+1} - \mathbf{e}_2\|/\|\mathbf{w}_n - \mathbf{e}_2\|$ remains stable at a level of approximately 2.05×10^{-4} for the first 4 iterations. This implies very fast linear convergence. However, as n increases further, it seems that the computer considers the sequence as "already converged", since $\|\mathbf{w}_{n+1} - \mathbf{e}_2\|/\|\mathbf{w}_n - \mathbf{e}_2\|$ becomes stable at 1.

3.4.2 The radius of convergence of FastICA with generalized Gaussian distribution

The aim of this section is to study the radius of convergence of the source signals. This notion is defined as follows.

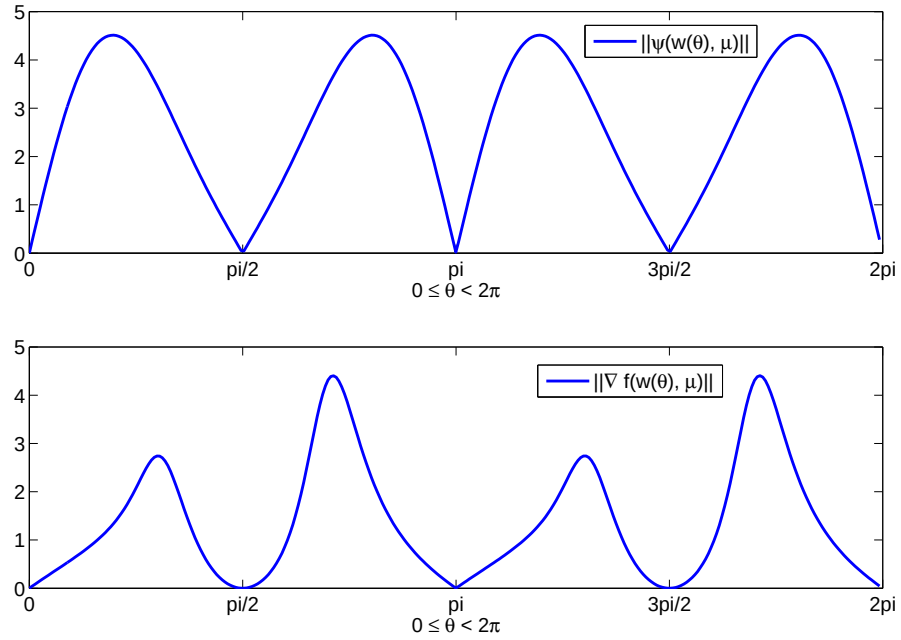

Figure 3.3: $\|\boldsymbol{\varphi}(\mathbf{w}(\theta),\mu)\|$ and $\|\nabla \boldsymbol{f}(\mathbf{w}(\theta),\mu)\|$ with $G(x) = x^4$ (kurtosis).

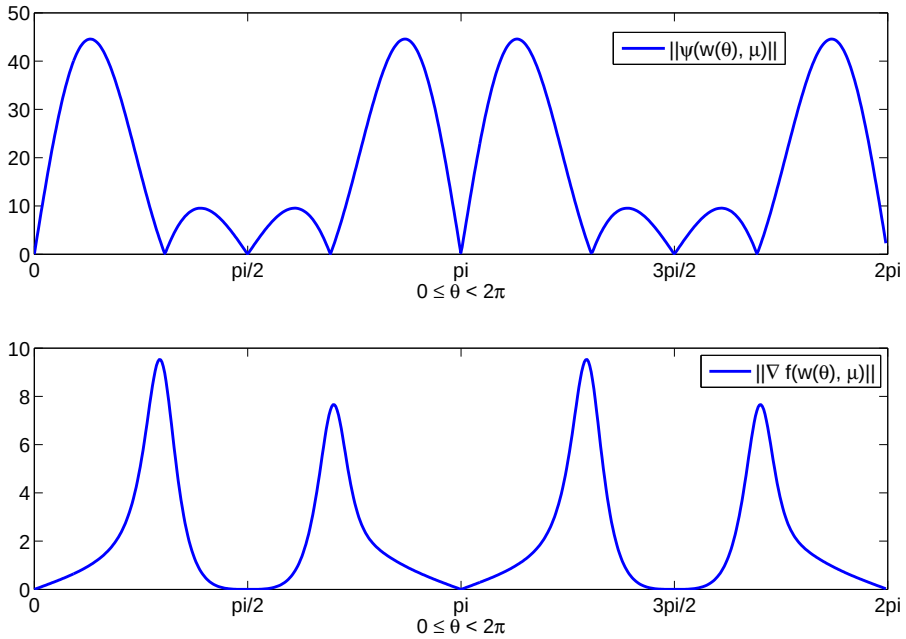

Figure 3.4: $\|\boldsymbol{\varphi}(\mathbf{w}(\theta),\mu)\|$ and $\|\nabla \boldsymbol{f}(\mathbf{w}(\theta),\mu)\|$ with $G(x) = x^6 - 15x^4 + 45x^2 - 15$ (Hermite polynomial $\mathcal{H}_6$).

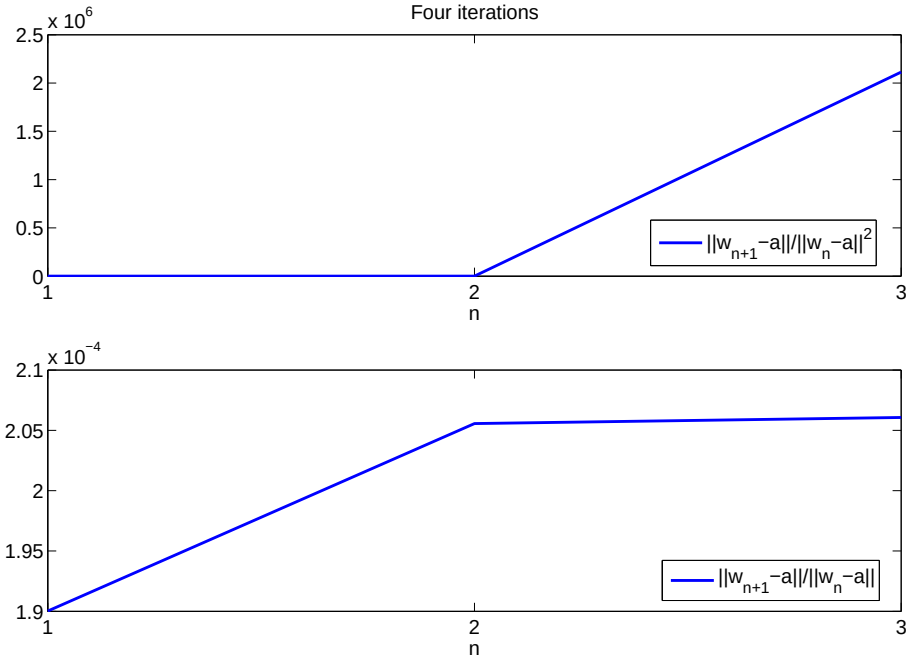

Figure 3.5: The FastICA algorithm is halted after four iterations.

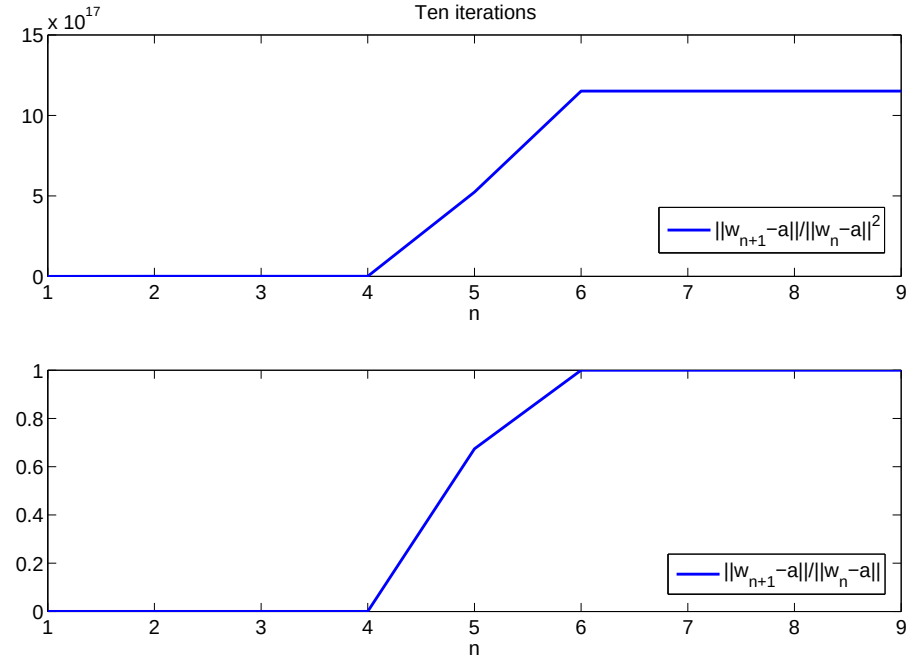

Figure 3.6: The FastICA algorithm is halted after ten iterations.

Definition 3.4.5. *The radius of convergence of s_i is the largest real number r, such that if the initial input $\mathbf{w}_0$ of the FastICA algorithm lies in the ball $\mathscr{B}_r(\mathbf{a}_i)$, then the FastICA algorithm is guaranteed to converge to $\mathbf{a}_i$.*

We have already encountered the radius of convergence in the previous sections, although the name was not formally employed. In Section 3.3, we actually studied the theoretical existence of the radius of convergence, while its quantitative value was still untouched. The latter is of great importance in practice, since it determines the likelihood of extracting of a source when the initial input $\mathbf{w}_0$ was chosen arbitrarily on $\mathcal{S}$. It is clear that the radius of convergence depends on the number of sources, the distribution of sources and the choice of nonlinearity function. In the simplest case that $d = 2$, each vector on $\mathcal{S}$ is parameterized by its angle θ on the unit circle. Thus the radius of convergence associated to s_i can be characterized by an interval $(\theta_i - \theta', \theta_i + \theta')$, where θ_i stands for the angle of $\mathbf{a}_i$. Fig 3.2 gives an example of the radius of convergence of s_2. If the angle θ_0 of the initial input $\mathbf{w}_0$ lies in $(1.005, 2.137)$ or $\left(\frac{\pi}{2} - 0.566, \frac{\pi}{2} + 0.566\right)$, then the FastICA algorithm converges to $\mathbf{e}_2$.

In this section, we will consider the source signal that has the generalized Gaussian distribution. The generalized Gaussian distribution is a parametric family of symmetrical distributions, whose probability density function (PDF) is given by (Tichavsky et al., 2006; Waheed & Salam, 2002)

$$f(x) = \frac{\lambda\beta}{2\Gamma(1/\lambda)} \exp(-(\beta|x|)^\lambda),$$

where λ and β are parameters and $\Gamma(\cdot)$ is the Gamma function. By the hypothesis of ICA, the source signal has unit variance. This is achieved by setting

$$\beta = \sqrt{\frac{\Gamma(3/\lambda)}{\Gamma(1/\lambda)}}.$$

The generalized Gaussian family encompasses the ordinary normal distribution for $\lambda = 2$, the Laplace distribution for $\lambda = 1$ and the uniform distribution in the limit $\lambda \to \infty$. In the sequel, the generalized Gaussian family with parameter λ will be abbreviated as $GG(\lambda)$.

Example 3.4.6. We suppose that s_1 has the generalized Gaussian distribution with parameter λ varying from 1 to 9 and s_2 has uniform distribution. For each $\lambda > 0$, there exists an interval $\big(-\theta(\lambda), \theta(\lambda)\big)$ that stands for the ball of convergence of the source signal s_1. In Fig 3.8, we plotted the curve $\lambda \to \theta(\lambda)$. From the figure, we observe that (i), for λ close to 2, i.e. when the signal s_1 is close to Gaussian, the angle $\theta(\lambda)$ tends to 0; (ii), for large λ, i.e. when the signal s_1 is close to uniform, the angle $\theta(\lambda)$ tends to $\pi/4$. Observation (i) means that if we choose arbitrarily $\mathbf{w}_0 \in \mathcal{S}$, then the FastICA

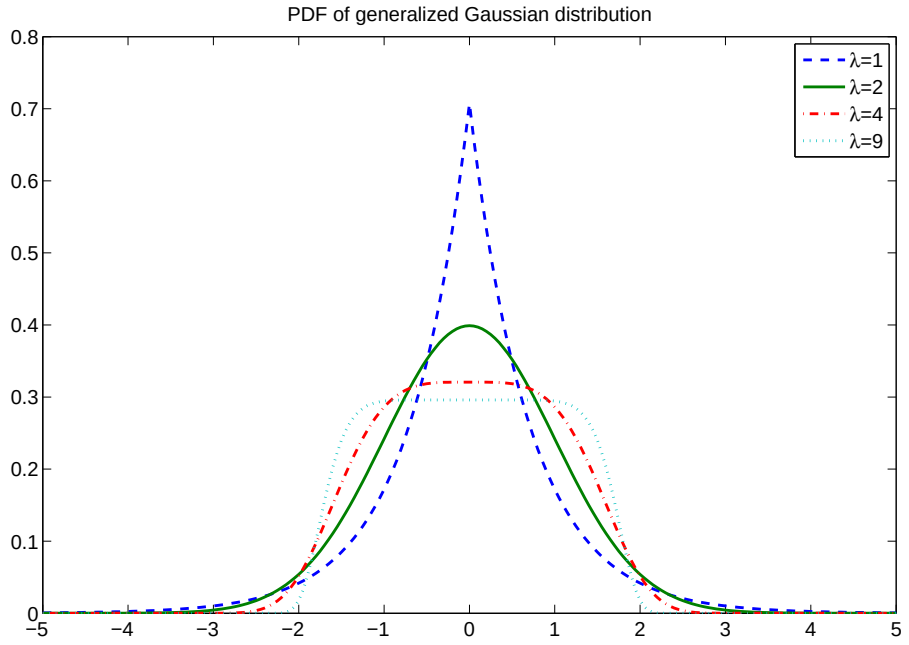

Figure 3.7: The probability density function of $GG(\lambda)$ for $\lambda = 1, 2, 4, 9$.

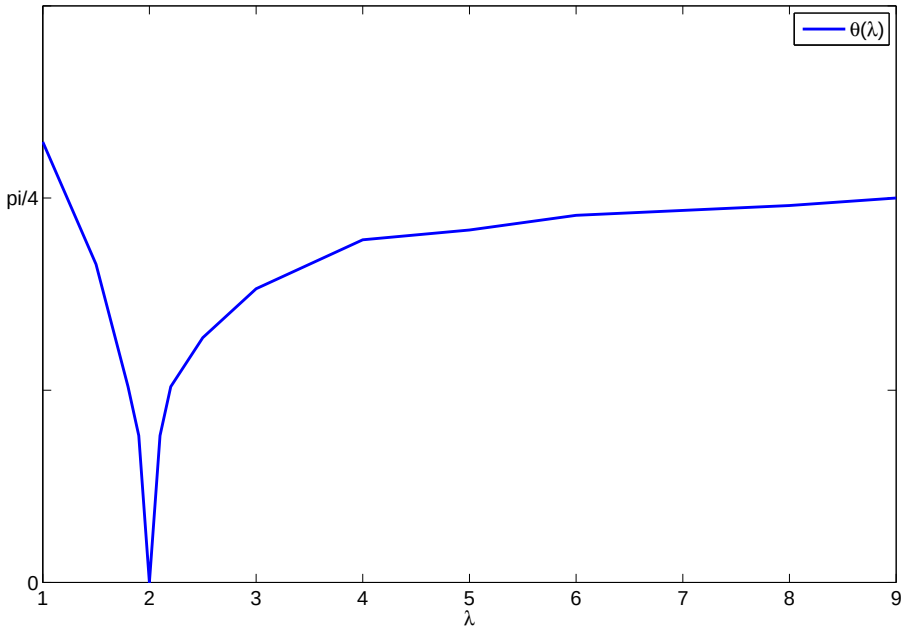

Figure 3.8: The radius of convergence of s_2, represented by an angle $\theta(\lambda)$, versus the parameter λ.

algorithm will very likely yield a sequence that converges to $\mathbf{e}_2$. In the extreme case that the signal $\lambda = 2$, we have $\theta(\lambda) = 0$ and hence we cannot extract s_1 directly using FastICA algorithm[1]. Observation (ii) is logical, since when the two sources have the same distribution, neither of the two should be "privileged" in the extraction.

3.5 Proofs

3.5.1 Proof of Proposition 3.2.1

The following lemmas are useful in the proof of Proposition 3.2.1.

Lemma 3.5.1. *For any* $\mathbf{w}, \mathbf{u} \in \mathcal{S}$*, we have*

$$(\mathbf{w}-\mathbf{u})^{\mathsf{T}}\mathbf{u} = -\frac{\|\mathbf{w}-\mathbf{u}\|^2}{2}.$$

Proof. We have

$$\|\mathbf{w}-\mathbf{u}\|^2 = (\mathbf{w}-\mathbf{u})^{\mathsf{T}}(\mathbf{w}-\mathbf{u}) = \mathbf{w}^{\mathsf{T}}\mathbf{w} - \mathbf{w}^{\mathsf{T}}\mathbf{u} - \mathbf{u}^{\mathsf{T}}\mathbf{w} + \mathbf{u}^{\mathsf{T}}\mathbf{u}.$$

Since $\mathbf{w}, \mathbf{u} \in \mathcal{S}$, there holds $\mathbf{w}^{\mathsf{T}}\mathbf{w} = \mathbf{u}^{\mathsf{T}}\mathbf{u} = 1$. It follows that

$$\|\mathbf{w}-\mathbf{u}\|^2 = 2(\mathbf{u}^{\mathsf{T}}\mathbf{u} - \mathbf{w}^{\mathsf{T}}\mathbf{u}) = -2(\mathbf{w}-\mathbf{u})^{\mathsf{T}}\mathbf{u}.$$

□

Lemma 3.5.2. *For any* $\mathbf{v} \in \mathcal{S}$*, we have*

$$\begin{aligned}&(\mathbf{w}-\mathbf{v})^{\mathbf{T}}\mathbb{E}[g'(\mathbf{v}^{\mathsf{T}}\mathbf{x})\mathbf{x}\mathbf{x}^{\mathsf{T}}](\mathbf{w}-\mathbf{v})\\ &= (\mathbf{w}-\mathbf{v})^{\mathsf{T}}\Big(\mathbb{E}[g'(\mathbf{v}^{\mathsf{T}}\mathbf{x})\mathbf{I} + \mathbf{L}(\mathbf{v},\mu)]\Big)(\mathbf{w}-\mathbf{v}) + \mathcal{O}(\|\mathbf{w}-\mathbf{v}\|^4),\end{aligned}$$

where

$$\mathbf{L}(\mathbf{v},\mu) \overset{\text{def}}{=} \mathbb{E}[g'(\mathbf{v}^{\mathsf{T}}\mathbf{x})\Pi_{\mathbf{v}}^{\perp}(\mathbf{x}\mathbf{x}^{\mathsf{T}} - \mathbf{I})\Pi_{\mathbf{v}}^{\perp}].$$

In particular, if $\mathbf{v} = \mathbf{a}$*, we have* $\mathbf{L}(\mathbf{a},\mu) = 0$.

Proof. Using (3.1.5), we have

$$\begin{aligned}\mathbb{E}[g'(\mathbf{v}^{\mathsf{T}}\mathbf{x})\mathbf{x}\mathbf{x}^{\mathsf{T}}] &= \mathbb{E}[g'(\mathbf{v}^{\mathsf{T}}\mathbf{x})(\Pi_{\mathbf{v}}\mathbf{x} + \Pi_{\mathbf{v}}^{\perp}\mathbf{x})(\Pi_{\mathbf{v}}\mathbf{x} + \Pi_{\mathbf{v}}^{\perp}\mathbf{x})^{\mathsf{T}}]\\ &= \mathbb{E}[g'(\mathbf{v}^{\mathsf{T}}\mathbf{x})(\Pi_{\mathbf{v}}\mathbf{x}\mathbf{x}^{\mathsf{T}}\Pi_{\mathbf{v}} + \Pi_{\mathbf{v}}^{\perp}\mathbf{x}\mathbf{x}^{\mathsf{T}}\Pi_{\mathbf{v}} + \Pi_{\mathbf{v}}\mathbf{x}\mathbf{x}^{\mathsf{T}}\Pi_{\mathbf{v}}^{\perp} + \Pi_{\mathbf{v}}^{\perp}\mathbf{x}\mathbf{x}^{\mathsf{T}}\Pi_{\mathbf{v}}^{\perp})].\end{aligned}$$

Note that

$$\begin{aligned}\mathbf{L}(\mathbf{v},\mu) &= \mathbb{E}[g'(\mathbf{v}^{\mathsf{T}}\mathbf{x})\Pi_{\mathbf{v}}^{\perp}(\mathbf{x}\mathbf{x}^{\mathsf{T}} - \mathbf{I})\Pi_{\mathbf{v}}^{\perp}]\\ &= \mathbb{E}[g'(\mathbf{v}^{\mathsf{T}}\mathbf{x})(\Pi_{\mathbf{v}}^{\perp}\mathbf{x}\mathbf{x}^{\mathsf{T}}\Pi_{\mathbf{v}} + \Pi_{\mathbf{v}}\mathbf{x}\mathbf{x}^{\mathsf{T}}\Pi_{\mathbf{v}}^{\perp} + \Pi_{\mathbf{v}}^{\perp}\mathbf{x}\mathbf{x}^{\mathsf{T}}\Pi_{\mathbf{v}}^{\perp} - \mathbf{I} + \mathbf{v}\mathbf{v}^{\mathsf{T}})].\end{aligned}$$

[1] In this case, however, we can extract s_2 and then use the deflation method to get s_1.

Hence we have

$$\mathbb{E}[g'(\mathbf{v}^\mathsf{T}\mathbf{x})\mathbf{x}\mathbf{x}^\mathsf{T}] = \mathbb{E}\big[g'(\mathbf{v}^\mathsf{T}\mathbf{x})\big(\mathbf{v}\mathbf{v}^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\mathbf{v}\mathbf{v}^\mathsf{T} - \mathbf{v}\mathbf{v}^\mathsf{T}\big)\big] + \mathbb{E}\big[g'(\mathbf{v}^\mathsf{T}\mathbf{x})\mathbf{I}\big] + \mathbf{L}(\mathbf{v},\mu). \tag{3.5.1}$$

By Lemma 3.5.1, we have:

$$(\mathbf{w}-\mathbf{v})^\mathsf{T}\mathbb{E}\big[g'(\mathbf{v}^\mathsf{T}\mathbf{x})\mathbf{v}\mathbf{v}^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\mathbf{v}\mathbf{v}^\mathsf{T}\big](\mathbf{w}-\mathbf{v}) = \mathbb{E}\big[(\mathbf{v}^\mathsf{T}\mathbf{x})^2 g'(\mathbf{v}^\mathsf{T}\mathbf{x})\big]\frac{\|\mathbf{w}-\mathbf{v}\|^4}{4},$$
$$(\mathbf{w}-\mathbf{v})^\mathsf{T}\mathbb{E}[g'(\mathbf{v}^\mathsf{T}\mathbf{x})\mathbf{v}\mathbf{v}^\mathsf{T}](\mathbf{w}-\mathbf{v}) = \mathbb{E}[g'(\mathbf{v}^\mathsf{T}\mathbf{x})]\frac{\|\mathbf{w}-\mathbf{v}\|^4}{4}.$$

It follows from (3.5.1) that

$$\begin{aligned}&(\mathbf{w}-\mathbf{v})^\mathsf{T}\mathbb{E}[g'(\mathbf{v}^\mathsf{T}\mathbf{x})\mathbf{x}\mathbf{x}^\mathsf{T}](\mathbf{w}-\mathbf{v})\\ &= (\mathbf{w}-\mathbf{v})^\mathsf{T}\Big(\mathbb{E}[g'(\mathbf{v}^\mathsf{T}\mathbf{x})]\mathbf{I} + \mathbf{L}(\mathbf{v},\mu)\Big)(\mathbf{w}-\mathbf{v}) + \mathbb{E}\big[((\mathbf{v}^\mathsf{T}\mathbf{x})^2-1)g'(\mathbf{v}^\mathsf{T}\mathbf{x})\big]\frac{\|\mathbf{w}-\mathbf{v}\|^4}{4}.\end{aligned}$$

□

Proof of Proposition 3.2.1. Using Taylor's formula, we have for any $a, b \in \mathbb{R}$:

$$G(b) = G(a) + g(a)(b-a) + \frac{1}{2}g'(a)(b-a)^2 + \frac{1}{6}g''\big(\delta b + (1-\delta)a\big)(b-a)^3,$$

where $0 < \delta < 1$. Now setting $b = \mathbf{w}^\mathsf{T}\mathbf{x}$, $a = \mathbf{v}^\mathsf{T}\mathbf{x}$ and taking mathematical expectation, we get

$$\begin{aligned}\mathbb{E}[G(\mathbf{w}^\mathsf{T}\mathbf{x})] = \mathbb{E}[G(\mathbf{v}^\mathsf{T}\mathbf{x})] + (\mathbf{w}-\mathbf{v})^\mathsf{T}\mathbb{E}[g(\mathbf{v}^\mathsf{T}\mathbf{x})\mathbf{x}] + \frac{1}{2}(\mathbf{w}-\mathbf{v})^\mathsf{T}\mathbb{E}[g'(\mathbf{v}^\mathsf{T}\mathbf{x})\mathbf{x}\mathbf{x}^\mathsf{T}](\mathbf{w}-\mathbf{v})\\ + \frac{1}{6}\mathbb{E}[g''(\boldsymbol{\xi}^\mathsf{T}\mathbf{x})((\mathbf{w}-\mathbf{v})^\mathsf{T}\mathbf{x})^3],\end{aligned}$$

where $\boldsymbol{\xi} = \delta\mathbf{w} + (1-\delta)\mathbf{v}$. Using the decomposition

$$\mathbf{x} = \Pi_\mathbf{v}\mathbf{x} + \Pi_\mathbf{v}^\perp\mathbf{x} = (\mathbf{v}^\mathsf{T}\mathbf{x})\mathbf{v} + (\mathbf{I} - \mathbf{v}\mathbf{v}^\mathsf{T})\mathbf{x},$$

we get

$$\begin{aligned}(\mathbf{w}-\mathbf{v})^\mathsf{T}\mathbb{E}[g(\mathbf{v}^\mathsf{T}\mathbf{x})\mathbf{x}] &= (\mathbf{w}-\mathbf{v})^\mathsf{T}\mathbb{E}[g(\mathbf{v}^\mathsf{T}\mathbf{x})(\Pi_\mathbf{v}\mathbf{x} + \Pi_\mathbf{v}^\perp\mathbf{x})]\\ &= \mathbb{E}[\mathbf{v}^\mathsf{T}\mathbf{x}g(\mathbf{v}^\mathsf{T}\mathbf{x})](\mathbf{w}-\mathbf{v})^\mathsf{T}\mathbf{v} + (\mathbf{w}-\mathbf{v})^\mathsf{T}\boldsymbol{\varphi}(\mathbf{v},\mu),\end{aligned}$$

where $\boldsymbol{\varphi}(\mathbf{v},\mu) \overset{\text{def}}{=} \mathbb{E}[g(\mathbf{v}^\mathsf{T}\mathbf{x})\Pi_\mathbf{v}^\perp\mathbf{x}]$. By Lemma 3.5.1, we have

$$\mathbb{E}[\mathbf{v}^\mathsf{T}\mathbf{x}g(\mathbf{v}^\mathsf{T}\mathbf{x})](\mathbf{w}-\mathbf{v})^\mathsf{T}\mathbf{v} = -\frac{\|\mathbf{w}-\mathbf{v}\|^2}{2}\mathbb{E}[\mathbf{v}^\mathsf{T}\mathbf{x}g(\mathbf{v}^\mathsf{T}\mathbf{x})].$$

It follows that

$$\begin{aligned}(\mathbf{w}-\mathbf{v})^\mathsf{T}\mathbb{E}[g(\mathbf{v}^\mathsf{T}\mathbf{x})\mathbf{x}] &= -\frac{\|\mathbf{w}-\mathbf{v}\|^2}{2}\mathbb{E}[\mathbf{v}^\mathsf{T}\mathbf{x}g(\mathbf{v}^\mathsf{T}\mathbf{x})] + (\mathbf{w}-\mathbf{v})^\mathsf{T}\boldsymbol{\varphi}(\mathbf{v},\mu)\\ &= -\frac{1}{2}(\mathbf{w}-\mathbf{v})^\mathsf{T}\mathbb{E}[\mathbf{v}^\mathsf{T}\mathbf{x}g(\mathbf{v}^\mathsf{T}\mathbf{x})](\mathbf{w}-\mathbf{v}) + (\mathbf{w}-\mathbf{v})^\mathsf{T}\boldsymbol{\varphi}(\mathbf{v},\mu)\end{aligned} \tag{3.5.2}$$

.

Using (3.5.2) and Lemma 3.5.2, we get

$$\begin{aligned} G(\mathbf{w},\mu) &= G(\mathbf{v},\mu) + (\mathbf{w}-\mathbf{v})^{\mathsf{T}}\mathbb{E}[g(\mathbf{v}^{\mathsf{T}}\mathbf{x})\mathbf{x}] + \frac{1}{2}(\mathbf{w}-\mathbf{v})^{\mathsf{T}}\mathbb{E}[g'(\mathbf{v}^{\mathsf{T}}\mathbf{x})\mathbf{x}\mathbf{x}^{\mathsf{T}}](\mathbf{w}-\mathbf{v}) \\ &\quad + \frac{1}{6}\mathbb{E}[g''(\boldsymbol{\xi}^{\mathsf{T}}\mathbf{x})((\mathbf{w}-\mathbf{v})^{\mathsf{T}}\mathbf{x})^3] \\ &= G(\mathbf{v},\mu) + (\mathbf{w}-\mathbf{v})^{\mathsf{T}}\boldsymbol{\varphi}(\mathbf{v}) + \frac{1}{2}(\mathbf{w}-\mathbf{v})^{\mathsf{T}}\mathbf{K}(\mathbf{v},\mu)(\mathbf{w}-\mathbf{v}) + \mathcal{O}(\|\mathbf{w}-\mathbf{v}\|^3). \end{aligned}$$

□

3.5.2 Proof of Proposition 3.3.5

In the sequel, we consider the ith column of $\mathbf{A}$, i.e. $\mathbf{a} = \mathbf{a}_i$. Besides, we denote by $\mathbf{b}_j^{\mathsf{T}}$ the jth row of $\mathbf{A}$, and by $\mathbf{A}_{ji}$ the ith entry of $\mathbf{b}_j$. It is easy to see that the following relations hold:

$$\mathbf{A}\mathbf{e}_i = \mathbf{a}_i, \quad \mathbf{A}\mathbf{b}_j = \mathbf{e}_j.$$

The hypothesis $G(x) = x^4$ implies that $g(x) = 4x^3$, $g'(x) = 12x^2$ and $g''(x) = 24x$. As a result, we have $\mathbb{E}[g''(s_i)] = 0$ and $\mathbb{E}[g'(s_i)] = 12$.

To show the cubic convergence speed, it suffices to prove that $\nabla^2 \boldsymbol{f}(\mathbf{a}_i,\mu) = 0$, or equivalently, $\partial_{w_j}\nabla \boldsymbol{f}(\mathbf{a}_i,\mu) = 0$ for all $j = 1,\dots,d$. Note that

$$\nabla \boldsymbol{f}(\mathbf{w},\mu)\|\boldsymbol{h}(\mathbf{w},\mu)\|^3 = \big(\|\boldsymbol{h}(\mathbf{w},\mu)\|^2\mathbf{I} - \boldsymbol{h}(\mathbf{w},\mu)\boldsymbol{h}(\mathbf{w},\mu)^{\mathsf{T}}\big)\nabla \boldsymbol{h}(\mathbf{w},\mu).$$

It follows that

$$\begin{aligned} &\partial_{w_i}\nabla \boldsymbol{f}(\mathbf{w},\mu)\|\boldsymbol{h}(\mathbf{w},\mu)\|^3 + \nabla \boldsymbol{f}(\mathbf{w},\mu)\partial_{w_i}\|\boldsymbol{h}(\mathbf{w},\mu)\| \\ &= \partial_{w_i}\Big(\big(\|\boldsymbol{h}(\mathbf{w},\mu)\|^2\mathbf{I} - \boldsymbol{h}(\mathbf{w},\mu)\boldsymbol{h}(\mathbf{w},\mu)^{\mathsf{T}}\big)\nabla \boldsymbol{h}(\mathbf{w},\mu)\Big). \end{aligned}$$

Since $\|\boldsymbol{h}(\mathbf{a}_i,\mu)\|^3 \neq 0$ and $\partial_{w_i}\|\boldsymbol{h}(\mathbf{a}_i,\mu)\| = 0$, equality $\partial_{w_j}\nabla \boldsymbol{f}(\mathbf{a}_i,\mu) = 0$ holds if and only if

$$\partial_{w_j}\Big(\big(\|\boldsymbol{h}(\mathbf{a}_i,\mu)\|^2\mathbf{I} - \boldsymbol{h}(\mathbf{a}_i,\mu)\boldsymbol{h}(\mathbf{a}_i,\mu)^{\mathsf{T}}\big)\nabla \boldsymbol{h}(\mathbf{a}_i,\mu)\Big) = 0. \tag{3.5.3}$$

Next, we shall show that (3.5.3) indeed holds for $j = 1,\dots,d$.

Step 1. Let's first prove that

$$\Big(\partial_{w_j}\big(\|\boldsymbol{h}(\mathbf{a}_i,\mu)\|^2\mathbf{I} - \boldsymbol{h}(\mathbf{a}_i,\mu)\boldsymbol{h}(\mathbf{a}_i,\mu)^{\mathsf{T}}\big)\Big)\nabla \boldsymbol{h}(\mathbf{a}_i,\mu) = 0.$$

From Proposition 3.3.3, we get $\nabla \boldsymbol{h}(\mathbf{a}_i,\mu) = \gamma(\mathbf{a}_i,\mu)\mathbf{a}_i\mathbf{a}_i^{\mathsf{T}}$. Besides, we have

$$\begin{aligned} \partial_{w_j}\|\boldsymbol{h}(\mathbf{w},\mu)\|^2 &= 2\boldsymbol{h}(\mathbf{w},\mu)^{\mathsf{T}}\partial_{w_j}\boldsymbol{h}(\mathbf{w},\mu) \\ \partial_{w_j}\big(\boldsymbol{h}(\mathbf{w},\mu)\boldsymbol{h}(\mathbf{w},\mu)^{\mathsf{T}}\big) &= \big(\partial_{w_j}\boldsymbol{h}(\mathbf{w},\mu)\big)\boldsymbol{h}(\mathbf{w},\mu)^{\mathsf{T}} + \boldsymbol{h}(\mathbf{w},\mu)\partial_{w_j}\boldsymbol{h}(\mathbf{w},\mu)^{\mathsf{T}}. \end{aligned}$$

We then deduce that

$$\begin{aligned}
\partial_{w_j}\|\boldsymbol{h}(\mathbf{a}_i,\mu)\|^2 &= 2\mathbb{E}[g'(s_i)\mathbf{a}_i - g(s_i)\mathbf{x}]^\mathsf{T}\mathbb{E}[g''(s_i)\mathbf{a}_i x_j + g'(s_i)\mathbf{e}_j - g'(s_i)\mathbf{x}x_j] \\
&= 2\mathbb{E}[g'(s_i)\mathbf{e}_i - g(s_i)\mathbf{s}]^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbb{E}[g''(s_i)\mathbf{e}_i x_j + g'(s_i)\mathbf{b}_j - g'(s_i)\mathbf{s}x_j] \\
&= 2\mathbb{E}[g'(s_i)\mathbf{e}_i - g(s_i)\mathbf{s}]^\mathsf{T}\big(24\mathbf{A}_{ji}\mathbf{e}_i + 12\mathbf{b}_j - (12\mathbf{b}_j - 12\mathbf{A}_{ji}\mathbf{e}_i + 12\mathbf{A}_{ji}\mathbb{E}[s_i^4]\mathbf{e}_i)\big) \\
&= 2H(\mathbf{a}_i,\mu)\mathbf{e}_i^\mathsf{T}(12\kappa\mathbf{e}_i) \\
&= 24\kappa H(\mathbf{a}_i,\mu), \\
\partial_{w_j}\big(\boldsymbol{h}(\mathbf{a}_i,\mu)\boldsymbol{h}(\mathbf{a}_i,\mu)^\mathsf{T}\big) &= \mathbb{E}[g'(s_i)\mathbf{a}_i - g(s_i)\mathbf{x}]\mathbb{E}[g''(s_i)\mathbf{a}_i x_j + g'(s_i)\mathbf{e}_j - g'(s_i)\mathbf{x}x_j]^\mathsf{T} \\
&\quad + \mathbb{E}[g''(s_i)\mathbf{a}_i x_j + g'(s_i)\mathbf{e}_j - g'(s_i)\mathbf{x}x_j]\mathbb{E}[g'(s_i)\mathbf{a}_i - g(s_i)\mathbf{x}]^\mathsf{T} \\
&= \mathbf{a}_i H(\mathbf{a}_i,\mu)(12\kappa\mathbf{e}_i)^\mathsf{T}\mathbf{A}^\mathsf{T} + \mathbf{A}(12\kappa\mathbf{e}_i)H(\mathbf{a}_i,\mu)\mathbf{a}_i^\mathsf{T} \\
&= 24\kappa H(\mathbf{a}_i,\mu)\mathbf{a}_i\mathbf{a}_i^\mathsf{T},
\end{aligned}$$

where $\kappa \stackrel{\text{def}}{=} 3 - \mathbb{E}[s_i^4]$. It follows that

$$\begin{aligned}
&\Big(\partial_{w_j}\big(\|\boldsymbol{h}(\mathbf{w},\mu)\|^2\mathbf{I} - \boldsymbol{h}(\mathbf{w},\mu)\boldsymbol{h}(\mathbf{w},\mu)^\mathsf{T}\big)\Big)\nabla\boldsymbol{h}(\mathbf{w},\mu) \\
&= \Big(24\kappa H(\mathbf{a}_i,\mu)\mathbf{I} - 24\kappa H(\mathbf{a}_i,\mu)\mathbf{a}_i\mathbf{a}_i^\mathsf{T}\Big)\gamma(\mathbf{a}_i,\mu)\mathbf{a}_i\mathbf{a}_i^\mathsf{T} = 0.
\end{aligned}$$

Step 2. We now show that

$$\big(\|\boldsymbol{h}(\mathbf{a}_i,\mu)\|^2\mathbf{I} - \boldsymbol{h}(\mathbf{a}_i,\mu)\boldsymbol{h}(\mathbf{a}_i,\mu)^\mathsf{T}\big)\partial_{w_i}\nabla\boldsymbol{h}(\mathbf{a}_i,\mu) = 0.$$

We have

$$\begin{aligned}
\partial_{w_j}\nabla\boldsymbol{h}(\mathbf{w},\mu) &= \partial_{w_j}\mathbb{E}_\mu[g''(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{w}\mathbf{x}^\mathsf{T} + g'(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{I} - g'(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{x}\mathbf{x}^\mathsf{T}] \\
&= \mathbb{E}_\mu[g'''(\mathbf{w}^\mathsf{T}\mathbf{x})x_j\mathbf{w}\mathbf{x}^\mathsf{T} + g''(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{e}_j\mathbf{x}^\mathsf{T} + g''(\mathbf{w}^\mathsf{T}\mathbf{x})x_j\mathbf{I} - g''(\mathbf{w}^\mathsf{T}\mathbf{x})x_j\mathbf{x}\mathbf{x}^\mathsf{T}]
\end{aligned}$$

It follows that

$$\partial_{w_j}\nabla\boldsymbol{h}(\mathbf{a}_i,\mu) = \mathbb{E}_\mu[g'''(s_i)x_j\mathbf{a}_i\mathbf{x}^\mathsf{T} + g''(s_i)\mathbf{e}_j\mathbf{x}^\mathsf{T} + g''(s_i)x_j\mathbf{I} - g''(s_i)x_j\mathbf{x}\mathbf{x}^\mathsf{T}].$$

Applying the assumption that $g = G' = 4x^3$, we get

$$\begin{aligned}
\mathbb{E}_\mu[g'''(s_i)x_j\mathbf{a}_i\mathbf{x}^\mathsf{T}] &= 0 \\
\mathbb{E}[g''(s_i)\mathbf{e}_j\mathbf{x}^\mathsf{T}] &= \mathbf{e}_j\mathbb{E}[g''(s_i)\mathbf{s}^\mathsf{T}]\mathbf{A}^\mathsf{T} \\
&= \mathbf{e}_j\mathbb{E}[g''(s_i)s_i]\mathbf{e}_i^\mathsf{T}\mathbf{A}^\mathsf{T} \\
&= 24\mathbf{e}_j\mathbf{a}_i^\mathsf{T}, \\
\mathbb{E}[g''(s_i)x_j\mathbf{I}] &= \mathbb{E}[g''(s_i)\mathbf{A}_{ji}s_i]\mathbf{I} \\
&= 24\mathbf{A}_{ji}\mathbf{I}
\end{aligned}$$

and

$$\mathbb{E}[g''(s_i)x_j\mathbf{x}\mathbf{x}^\mathsf{T}] = \mathbf{A}\mathbb{E}[g''(s_i)x_j\mathbf{s}\mathbf{s}^\mathsf{T}]\mathbf{A}^\mathsf{T},$$

where

$$
\begin{aligned}
&\mathbb{E}[g''(s_i)x_j\mathbf{s}\mathbf{s}^\mathsf{T}] \\
&= \begin{pmatrix}
\mathbb{E}[g''(s_i)x_js_1s_1] & \cdot & \mathbb{E}[g''(s_i)x_js_1s_i] & \cdot & \mathbb{E}[g''(s_i)x_js_1s_d] \\
\vdots & \cdot & \vdots & \cdot & \vdots \\
\mathbb{E}[g''(s_i)x_js_is_1] & \cdot & \mathbb{E}[g''(s_i)x_js_is_i] & \cdot & \mathbb{E}[g''(s_i)x_js_is_d] \\
\vdots & \cdot & \vdots & \cdot & \vdots \\
\mathbb{E}[g''(s_i)x_js_ds_1] & \cdot & \mathbb{E}[g''(s_i)x_js_ds_i] & \cdot & \mathbb{E}[g''(s_i)x_js_ds_d]
\end{pmatrix} \\
&= \begin{pmatrix}
24\mathbf{A}_{ji} & \cdot & 24\mathbf{A}_{j1} & \cdot & 0 \\
\vdots & \cdot & \vdots & \cdot & \vdots \\
24\mathbf{A}_{j1} & \cdot & 24\mathbf{A}_{ji}\mathbb{E}[s_i^4] & \cdot & 24\mathbf{A}_{jd} \\
\vdots & \cdot & \vdots & \cdot & \vdots \\
0 & \cdot & 24\mathbf{A}_{jd} & \cdot & 24\mathbf{A}_{ji}
\end{pmatrix} \\
&= 24\mathbf{A}_{ji}\mathbf{I} + 24\mathbf{e}_i\mathbf{b}_j^\mathsf{T} + 24\mathbf{b}_j\mathbf{e}_i^\mathsf{T} - 72\mathbf{A}_{ji}\mathbf{e}_i\mathbf{e}_i^\mathsf{T} + 24\mathbf{A}_{ji}\mathbb{E}[s_i^4]\mathbf{e}_i\mathbf{e}_i^\mathsf{T}.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\mathbb{E}[g''(s_i)x_j\mathbf{x}\mathbf{x}^\mathsf{T}] &= \mathbf{A}(24\mathbf{A}_{ji}\mathbf{I} + 24\mathbf{e}_i\mathbf{b}_j^\mathsf{T} + 24\mathbf{b}_j\mathbf{e}_i^\mathsf{T} - 72\mathbf{A}_{ji}\mathbf{e}_i\mathbf{e}_i^\mathsf{T} + 24\mathbf{A}_{ji}\mathbb{E}[s_i^4]\mathbf{e}_i\mathbf{e}_i^\mathsf{T})\mathbf{A}^\mathsf{T} \\
&= 24\mathbf{A}_{ji}\mathbf{I} + 24\mathbf{a}_i\mathbf{e}_j^\mathsf{T} + 24\mathbf{e}_j\mathbf{a}_i^\mathsf{T} - 72\mathbf{A}_{ji}\mathbf{a}_i\mathbf{a}_i^\mathsf{T} + 24\mathbf{A}_{ji}\mathbb{E}[s_i^4]\mathbf{a}_i\mathbf{a}_i^\mathsf{T}.
\end{aligned}
$$

We then deduce that

$$
\begin{aligned}
\partial_{w_j}\nabla\boldsymbol{h}(\mathbf{a}_i,\mu) &= \mathbb{E}_\mu[g'''(s_i)x_j\mathbf{a}_i\mathbf{x}^\mathsf{T} + g''(s_i)\mathbf{e}_j\mathbf{x}^\mathsf{T} + g''(s_i)x_j\mathbf{I} - g''(s_i)x_j\mathbf{x}\mathbf{x}^\mathsf{T}] \\
&= 24\mathbf{e}_j\mathbf{a}_i^\mathsf{T} + 24\mathbf{A}_{ji}\mathbf{I} - (24\mathbf{A}_{ji}\mathbf{I} + 24\mathbf{a}_i\mathbf{e}_j^\mathsf{T} + 24\mathbf{e}_j\mathbf{a}_i^\mathsf{T} - 72\mathbf{A}_{ji}\mathbf{a}_i\mathbf{a}_i^\mathsf{T} \\
&\quad + 24\mathbf{A}_{ji}\mathbb{E}[s_i^4]\mathbf{a}_i\mathbf{a}_i^\mathsf{T}) \\
&= -24\mathbf{a}_i\mathbf{e}_j^\mathsf{T} + 72\mathbf{A}_{ji}\mathbf{a}_i\mathbf{a}_i^\mathsf{T} - 24\mathbf{A}_{ji}\mathbb{E}[s_i^4]\mathbf{a}_i\mathbf{a}_i^\mathsf{T} \\
&= -24\mathbf{a}_i\mathbf{e}_j^\mathsf{T} + 24\kappa\mathbf{A}_{ji}\mathbf{a}_i\mathbf{a}_i^\mathsf{T}.
\end{aligned}
$$

On the other hand, we get from Proposition 3.3.3 that

$$
\|\boldsymbol{h}(\mathbf{a}_i,\mu)\|^2\mathbf{I} - \boldsymbol{h}(\mathbf{a}_i,\mu)\boldsymbol{h}(\mathbf{a}_i,\mu)^\mathsf{T} = H(\mathbf{a}_i,\mu)^2(\mathbf{I} - \mathbf{a}_i\mathbf{a}_i^\mathsf{T}).
$$

As a result,

$$
\begin{aligned}
&\big(\|\boldsymbol{h}(\mathbf{a}_i,\mu)\|^2\mathbf{I} - \boldsymbol{h}(\mathbf{a}_i,\mu)\boldsymbol{h}(\mathbf{a}_i,\mu)^\mathsf{T}\big)\partial_{w_j}\nabla\boldsymbol{h}(\mathbf{a}_i,\mu) \\
&= H(\mathbf{a}_i,\mu)^2(\mathbf{I} - \mathbf{a}_i\mathbf{a}_i^\mathsf{T})(-24\mathbf{a}_i\mathbf{e}_j^\mathsf{T} + 24\kappa\mathbf{A}_{ji}\mathbf{a}_i\mathbf{a}_i^\mathsf{T}) = 0.
\end{aligned}
$$

Finally, using the fact that

$$
\begin{aligned}
&\partial_{w_j}\Big(\big(\|\boldsymbol{h}(\mathbf{w},\mu)\|^2\mathbf{I} - \boldsymbol{h}(\mathbf{w},\mu)\boldsymbol{h}(\mathbf{w},\mu)^\mathsf{T}\big)\nabla\boldsymbol{h}(\mathbf{w},\mu)\Big) \\
&= \partial_{w_j}\big(\|\boldsymbol{h}(\mathbf{w},\mu)\|^2\mathbf{I} - \boldsymbol{h}(\mathbf{w},\mu)\boldsymbol{h}(\mathbf{w},\mu)^\mathsf{T}\big)\nabla\boldsymbol{h}(\mathbf{w},\mu) \\
&\quad + \big(\|\boldsymbol{h}(\mathbf{w},\mu)\|^2\mathbf{I} - \boldsymbol{h}(\mathbf{w},\mu)\boldsymbol{h}(\mathbf{w},\mu)^\mathsf{T}\big)\partial_{w_i}\nabla\boldsymbol{h}(\mathbf{w},\mu),
\end{aligned}
$$

we achieve the proof.

Chapter 4

Four FastICA estimators

Contents

In this chapiter, we work with the empirical ICA model 2.2.1 with Assumption 3. Besides, we suppose that $\mathbb{E}[\mathbf{x}] = 0$ and $\mathrm{Cov}(\mathbf{x}) = \mathbf{I}$ so that μ_N^1-μ_N^4 can all be tackled in a unified framework. The aim of this chapter is to prove the convergence of the empirical FastICA algorithm and the consistency of the FastICA estimator with respect to μ_N^1-μ_N^4. We start by introducing the notion of empirical contrast function and the uniform strong law of large numbers (USLLN). By generalizing results established in Chapter 3, we establish a link between the local minimizers of the empirical contrast functions and the fixed points of the empirical FastICA functions. Finally, we show that the empirical FastICA algorithms is almost surely convergent provided that the sample size N is large enough.

4.1 Approach to empirical FastICA

One purpose of this thesis is to investigate the convergence of empirical FastICA algorithm. Intuitively, the convergence should take place with large probability for large N, since the empirical FastICA algorithm is merely an approximation of the theoretical one, and these two can be arbitrarily close. Although this conjecture is supported by numerical simulation, and is implicitly taken as a hypothesis by many authors, there does not yet exist a rigorous proof in the community. In this chapter, we aim at filling this blank. Note that since the empirical FastICA algorithm depends on the particular realizations of the observed signal, we cannot expect to get a deterministic result. Fig 4.1 illustrates an example where the empirical FastICA algorithm fails to converge (see Example 4.3.1 for more detail).

In Chapter 3, we used a fixed-point argument to prove the convergence of theoretical FastICA algorithm. Here, we hope to follow the same idea. Recall that the key to the convergence of theoretical FastICA algorithm was

Proposition 3.2.5 and 3.3.1. A careful examination of these results reveals that they are actually independent of the probability measure involved. That is to say, we can replace μ by μ_N^k in the statement of Proposition 3.2.5 and Proposition 3.3.1, while the conclusion still holds with exactly the same proof. This remark leads us to the notion of the *empirical contrast function*, which can be considered as a generalization of the contrast function defined in (2.1.13).

Definition 4.1.1. *Let $G(\cdot)$ be a twice continuously differentiable nonlinear and nonquadratic function, and μ_N^k be the probability measure defined in (2.2.4)-(2.2.7) for $k = 1, 2, 3, 4$. The function $G(\cdot, \mu_N^k) : \mathcal{S} \to \mathbb{R}$ defined by*

$$G(\mathbf{w}, \mu_N^k) \stackrel{\text{def}}{=} \mathbb{E}_{\mu_N^k}[G(\mathbf{w}^\mathsf{T}\mathbf{z})]$$

is called the **empirical contrast function** *with respect to measure μ_N^k, or simply the empirical contrast function.*

Remark 4.1.2. We recall that the operator $\mathbb{E}_{\mu_N^k}[\cdot]$ stands for an average, see (2.2.8). Then (4.1.1) can be explicitly written as

$$G(\mathbf{w}, \mu_N^1) = \frac{1}{N}\sum_{i=1}^{N} G(\mathbf{w}^\mathsf{T}\mathbf{x}(t)) \tag{4.1.1}$$

$$G(\mathbf{w}, \mu_N^2) = \frac{1}{N}\sum_{i=1}^{N} G(\mathbf{w}^\mathsf{T}(\mathbf{x}(t) - \bar{\mathbf{x}})) \tag{4.1.2}$$

$$G(\mathbf{w}, \mu_N^3) = \frac{1}{N}\sum_{i=1}^{N} G(\mathbf{w}^\mathsf{T}\mathbf{Q}_N^{-1/2}\mathbf{x}(t)) \tag{4.1.3}$$

$$G(\mathbf{w}, \mu_N^4) = \frac{1}{N}\sum_{i=1}^{N} G(\mathbf{w}^\mathsf{T}\mathbf{C}_N^{-1/2}(\mathbf{x}(t) - \bar{\mathbf{x}})). \tag{4.1.4}$$

All the results established in Section 3.2, with the exception of Proposition 3.2.6, can be generalized for the empirical contrast function and the empirical FastICA algorithm. Proposition 3.2.6 cannot be generalized because it relies on the fact that $\mathbf{a}^\mathsf{T}\mathbf{x}$ and $(\mathbf{a}\mathbf{a}^\mathsf{T} - \mathbf{I})\mathbf{x}$ are independent with respect to μ, thanks to the fundamental hypothesis of ICA (see Lemma 3.1.2); while for random vector $\mathbf{z}$ having probability distribution μ_N^k, in general, there does not necessarily exist a vector $\mathbf{v}$ such that $\mathbf{z}^\mathsf{T}\mathbf{v}$ and $(\mathbf{I} - \mathbf{v}\mathbf{v}^\mathsf{T})\mathbf{z}$ are independent with respect to μ_N^k.

Let's state the generalized version of Proposition 3.2.5 and Proposition 3.3.1.

Proposition 4.1.3. *If $\mathbf{v}$ is a local minimizer of $G(\cdot, \mu_N^k)$ on $\mathcal{S}$, and if $H(\mathbf{v}, \mu_N^k) > 0$, then it is a fixed point of $\boldsymbol{f}(\cdot, \mu_N^k)$.*

Proposition 4.1.4. *Let* $\mathbf{v}$ *be a fixed point of* $\boldsymbol{f}(\cdot, \mu_N^k)$. *If* $\|\nabla \boldsymbol{f}(\mathbf{v}, \mu_N^k)\| < 1$, *then starting near* $\mathbf{v}$, *the empirical FastICA algorithm converges to* $\mathbf{v}$.

Proposition 4.1.3 reveals the link between the local minimizer of $G(\cdot, \mu_N^k)$ and the fixed point of the empirical FastICA algorithm, while Proposition 3.3.1 gives a sufficient condition for the empirical FastICA algorithm to converge. Now we are halfway to our goal, and it remains to verify the condition $H(\mathbf{v}, \mu_N^k) > 0$ and $\|\nabla \boldsymbol{f}(\mathbf{v}, \mu_N^k)\| < 1$. Note that we have already $H(\mathbf{a}, \mu) > 0$ and $\|\nabla \boldsymbol{f}(\mathbf{a}, \mu)\| = 0$ by Proposition 3.3.3, hence the former conditions can be achieved through proving the convergence

$$H(\mathbf{v}, \mu_N^k) \xrightarrow[N\to\infty]{} H(\mathbf{a}, \mu_N^k), \quad \nabla \boldsymbol{f}(\mathbf{v}, \mu_N^k) \xrightarrow[N\to\infty]{} \nabla \boldsymbol{f}(\mathbf{a}, \mu). \quad (4.1.5)$$

Since the vector $\mathbf{v}$, being random itself, varies according to the sample and the sample size N, we would need a uniform type of convergence rather than a point-wise one to achieve 4.1.5. The key tool here is the Uniform Strong Law of Large Numbers (USLLN). The following version of USLLN can be found in (Bierens, 2005). For a detailed discussion of this theorem, we refer to (Newey, 1991; Andrews, 1992).

Theorem 4.1.5 (USLLN)**.** *Let* $\mathbf{x}(1), \ldots, \mathbf{x}(N)$ *be a random sample of a d-variate distribution, and let* $\boldsymbol{\theta}$ *be non random vectors in a compact subset* $\Theta \in \mathbb{R}^m$. *Moreover, let* $h(\mathbf{x}, \boldsymbol{\theta})$ *be a Borel measurable function on* $\mathbb{R}^d \times \Theta$ *such that for each* $\mathbf{x}$, $h(\mathbf{x}, \boldsymbol{\theta})$ *is a continuous function on* Θ. *Finally, assume that* $\mathbb{E}[\sup_{\boldsymbol{\theta}\in\Theta} |h(\mathbf{x}(1), \boldsymbol{\theta})|] < \infty$. *Then with probability one we have*

$$\lim_{N\to\infty} \sup_{\boldsymbol{\theta}\in\Theta} \left\| \frac{1}{N} \sum_{t=1}^{N} h\big(\mathbf{x}(t), \boldsymbol{\theta}\big) - \mathbb{E}\big[h(\mathbf{x}(1), \boldsymbol{\theta})\big] \right\| = 0.$$

Using USLLN, we can prove the following result:

Proposition 4.1.6. *For* $k = 1, 2, 3, 4$, *the following uniform convergence holds:*

$$\sup_{\mathbf{w}\in\mathcal{S}} \|G(\mathbf{w}, \mu_N^k) - G(\mathbf{w}, \mu)\| \xrightarrow[N\to\infty]{a.s.} 0 \quad (4.1.6)$$

$$\sup_{\mathbf{w}\in\mathcal{S}} \|\boldsymbol{h}(\mathbf{w}, \mu_N^k) - \boldsymbol{h}(\mathbf{w}, \mu)\| \xrightarrow[N\to\infty]{a.s.} 0 \quad (4.1.7)$$

$$\sup_{\mathbf{w}\in\mathcal{S}} \|\nabla \boldsymbol{h}(\mathbf{w}, \mu_N^k) - \nabla \boldsymbol{h}(\mathbf{w}, \mu)\| \xrightarrow[N\to\infty]{a.s.} 0. \quad (4.1.8)$$

$$\sup_{\mathbf{w}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}} \|\boldsymbol{f}(\mathbf{w}, \mu_N^k) - \boldsymbol{f}(\mathbf{w}, \mu)\| \xrightarrow[N\to\infty]{a.s.} 0 \quad (4.1.9)$$

$$\sup_{\mathbf{w}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}} \|\nabla \boldsymbol{f}(\mathbf{w}, \mu_N^k) - \nabla \boldsymbol{f}(\mathbf{w}, \mu)\| \xrightarrow[N\to\infty]{a.s.} 0. \quad (4.1.10)$$

Proof. See Section 4.4. □

4.2 Local convergence of empirical FastICA algorithm

Proposition 4.2.1. *For any $r > 0$, the empirical contrast function $G(\cdot, \mu_N^k)$ has a local minimizer in $\mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}$ with probability one for large N.*

Proof. First, we note that $G(\mathbf{w}, \mu_N^k)$ has a minimizer on the compact $\mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}$ since $G(\cdot, \mu_N^k)$ is a continuous function. Now we prove that this minimizer is located inside of $\mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}$.

Since $\mathbf{a}$ is the unique local minimizer of $G(\cdot, \mu)$ on $\mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}$, the value of $G(\cdot, \mu)$ on the frontier of $\mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}$ is strictly larger than $G(\mathbf{a}, \mu)$. Denote $\epsilon \stackrel{\text{def}}{=} \inf_{\mathbf{w} \in \mathcal{S}, \|\mathbf{w}-\mathbf{a}\|=r} G(\mathbf{w}, \mu)$. By the uniform convergence

$$\sup_{\mathbf{w} \in \mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}} \|G(\mathbf{w}, \mu_N^k) - G(\mathbf{w}, \mu)\| \xrightarrow[N \to \infty]{a.s.} 0,$$

we obtain that with probability one, there exists N, such that

$$\sup_{\mathbf{w} \in \mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}} \|G(\mathbf{w}, \mu_N^k) - G(\mathbf{w}, \mu)\| < \frac{\epsilon}{2}.$$

It follows that

$$\inf_{\mathbf{w} \in \mathcal{S}, \|\mathbf{w}-\mathbf{a}\|=r} G(\mathbf{w}, \mu_N^k) > G(\mathbf{a}, \mu_N^k).$$

This means that the local minimizer of $G(\cdot, \mu_N^k)$ is inside of $\mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}$. □

Lemma 4.2.2. *There exists $r > 0$, such that with probability one there holds $\inf_{\mathbf{w} \in \mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}} H(\mathbf{w}, \mu_N^k) > 0$ for large N.*

Proof. First, we note that $H(\mathbf{w}, \nu) = \mathbf{w}^\mathsf{T} \boldsymbol{h}(\mathbf{w}, \nu)$ for any $\mathbf{w} \in \mathcal{S}$. Then from (4.1.7) we deduce immediately

$$\sup_{\mathbf{w} \in \mathcal{S}} \|H(\mathbf{w}, \mu_N^k) - H(\mathbf{w}, \mu)\| \xrightarrow[N \to \infty]{a.s.} 0. \tag{4.2.1}$$

Besides, since $H(\mathbf{a}, \mu) > 0$ and $H(\cdot, \mu)$ is a continuous function, there exists $r > 0$ such that

$$\inf_{\mathbf{w} \in \mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}} H(\mathbf{w}, \mu_N^k) > 0. \tag{4.2.2}$$

By the triangle inequality, we have $H(\mathbf{w}, \mu_N^k) \geq H(\mathbf{w}, \mu) - |H(\mathbf{w}, \mu) - H(\mathbf{w}, \mu_N^k)|$. Then it follows that

$$\inf_{\mathbf{w} \in \mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}} H(\mathbf{w}, \mu_N^k) \geq \inf_{\mathbf{w} \in \mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}} H(\mathbf{w}, \mu) - \sup_{\mathbf{w} \in \mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}} |H(\mathbf{w}, \mu) - H(\mathbf{w}, \mu_N^k)|.$$

Applying uniform convergence (4.2.1) and bound (4.2.2), we then get the conclusion. □

Proposition 4.2.3. *There exists $0 < K < 1$, $r > 0$ such that with probability one, we have* $\sup_{\mathbf{w}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}} \|\nabla \boldsymbol{f}(\mathbf{w}, \mu_N^k)\| \leq K$ *for large N.*

Proof. Since $\|\boldsymbol{h}(\mathbf{w}, \mu)\| > 0$ holds in a neighborhood of $\mathbf{a}$, the function $\boldsymbol{f}(\cdot, \mu)$ is continuous in this neighborhood. Therefore, by the fact $\nabla \boldsymbol{f}(\mathbf{a}, \mu) = 0$, there exists $r > 0$ such that in $\mathbf{w} \in \mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}$, we have $\sup_{\mathbf{w}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}} \|\nabla \boldsymbol{f}(\mathbf{w}, \mu)\| < K$. It follows that

$$\begin{aligned}
&\sup_{\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}} \|\nabla \boldsymbol{f}(\mathbf{w}, \mu_N^4)\| \\
\leq &\sup_{\mathbf{w}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}} \|\nabla \boldsymbol{f}(\mathbf{w}, \mu_N^4) - \nabla \boldsymbol{f}(\mathbf{w}, \mu)\| + \sup_{\mathbf{w}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}} \|\nabla \boldsymbol{f}(\mathbf{w}, \mu)\| \\
\leq &\sup_{\mathbf{w}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}} \|\nabla \boldsymbol{f}(\mathbf{w}, \mu_N^4) - \nabla \boldsymbol{f}(\mathbf{w}, \mu)\| + K.
\end{aligned}$$

Then applying (4.1.9) of Proposition 4.1.6, we achieve the proof. □

Theorem 4.2.4. *There exists $r > 0$ such that if $\mathbf{w}_0 \in \mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}$, then with probability one, the empirical FastICA algorithm with respect to measure μ_N^k converges to a local minimizer $\mathbf{a}_N^k$ of the contrast function $G(\mathbf{w}, \mu_N^k)$ for large N. Moreover, $\mathbf{a}_N^k$ is consistent estimator of $\mathbf{a}$.*

Proof. By Theorem 4.2.1 and Lemma 4.2.2, we deduce that there exists a positive number r, which can be arbitrarily small, such that with probability one $G(\cdot, \mu_N^k)$ has a local minimizer $\mathbf{a}_N^k$ in $\mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}$ and $H(\mathbf{a}_N^k, \mu_N^k) > 0$ for large N. Then it follows from Proposition 4.1.3 that $\mathbf{a}_N^k$ is a fixed point of the empirical FastICA function $\boldsymbol{f}(\cdot, \mu_N^k)$. Besides, by Proposition 4.2.3 we have also $\|\nabla \boldsymbol{f}(\mathbf{a}_N^k, \mu_N^k)\| < 1$. In view of these facts, applying Proposition 4.1.4, we achieve the convergence of empirical FastICA algorithm. Lastly, the consistency of the estimator $\mathbf{a}_N^k$ comes from Theorem 4.2.1. □

Remark 4.2.5. Theorem 4.2.4 does not say that the limit of empirical FastICA is *necessarily* a local minimizer of the empirical contrast function, it states only that this event occurs with probability one for large N. In practice, especially when the sample size is relatively small, it is possible, although very rare, that FastICA eventually converges to a saddle point of the empirical contrast function. Some authors (Tichavsky et al., 2006; Oja & Yuan, 2006) have noticed the fact that the empirical FastICA algorithm searches the stationary point of the empirical contrast function, but none has given a proof.

4.3 Numerical results

In this section, all the settings are the same as in Chapter 2, i.e. we consider the case $d = 2$ that the two source signals s_1, s_2 have respectively Laplace

	N=10	N=100	N=200	N=500	N=1000
kurt, $k=1$	1743	**204**	84	10	1
kurt, $k=2$	**1443**	226	77	7	**0**
kurt, $k=3$	3921	538	40	**0**	**0**
kurt, $k=4$	3883	571	**36**	**0**	**0**
gaus, $k=1$	978	241	58	2	**0**
gaus, $k=2$	1061	229	51	1	**0**
gaus, $k=3$	951	**0**	**0**	**0**	**0**
gaus, $k=4$	**732**	**0**	**0**	**0**	**0**
tanh, $k=1$	850	827	480	126	11
tanh, $k=2$	906	865	539	134	10
tanh, $k=3$	203	**0**	0	**0**	**0**
tanh, $k=4$	**56**	**0**	0	**0**	**0**

Table 4.1: Number of failure of convergence among 10000 trials.

and uniform distribution. We take $G(x) = x^4$ and suppose that $\mathbf{A} = \mathbf{I}$. Then we have $\mathbf{w}(\theta) = (\cos(\theta), \sin(\theta))^\mathsf{T}$, and

$$\theta \to G(\mathbf{w}(\theta), \nu) = \mathbb{E}[G(\cos(\theta)s_1 + \sin(\theta)s_2)]. \tag{4.3.1}$$

We recall that $\mathbf{A} = (\mathbf{e}_1, \mathbf{e}_2)$ and $\mathbf{w}(0) = \mathbf{e}_1$, $\mathbf{w}(\pi/2) = \mathbf{e}_2$, $\mathbf{w}(\pi) = -\mathbf{e}_1$, $\mathbf{w}(3\pi/2) = -\mathbf{e}_2$.

Example 4.3.1. Fig. 4.1 illustrates a failure of convergence of the empirical FastICA algorithm. In the simulation, we generate a very small sample with $N = 10$, run FastICA with respect to μ_N^1 for 50 iterations and plot the quantity $\epsilon_n \overset{\text{def}}{=} \|\mathbf{w}_{n+1} - \mathbf{w}_n\|$ versus the iteration times n. Clearly, if the algorithm converges, we should have $\epsilon_n \to 0$. The simulation shows that both cases are possible: depending on the specific sample, the FastICA algorithm can be either convergent or non-convergent. Fig 4.2 shows a case of successful convergence under the same settings. In fact, for the case of $N = 10$ and kurtosis nonlinearity function, the convergence of empirical FastICA with respect to μ_N^1 takes place more than 80% of the times.

Tab. 4.1 shows the number of failure of convergence among 10000 independent trials. Different sample sizes (from $N = 10$ to $N = 1000$), different nonlinearity functions ("kurtosis", "Gauss" and "tanh") and different measures μ_N^k $(k = 1, 2, 3, 4)$ are considered. Numbers in boldface signify that they correspond to the lowest failure rate in their category. From the table, we observe that: 1) In most cases, the centering and whitening procedure can significantly improve the chance of convergence. If the sample size larger than 500, then the convergence of the algorithm with respect to μ_N^3 or μ_N^4 is almost guaranteed for all three nonlinearity functions. The only exception is case of kurtosis nonlinearity along with a small sample size ($N \leq 100$). In

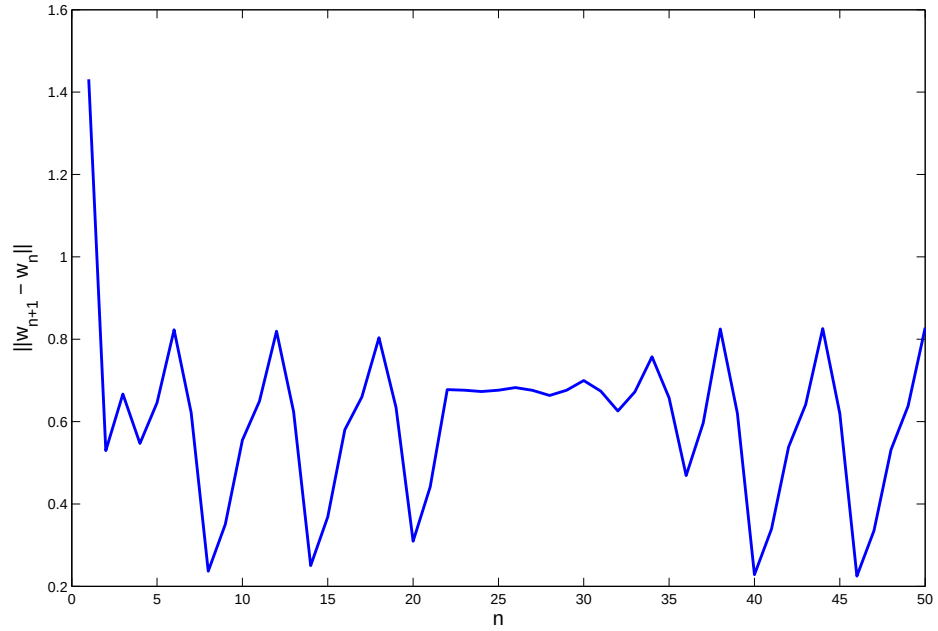

Figure 4.1: FastICA fails to converge.

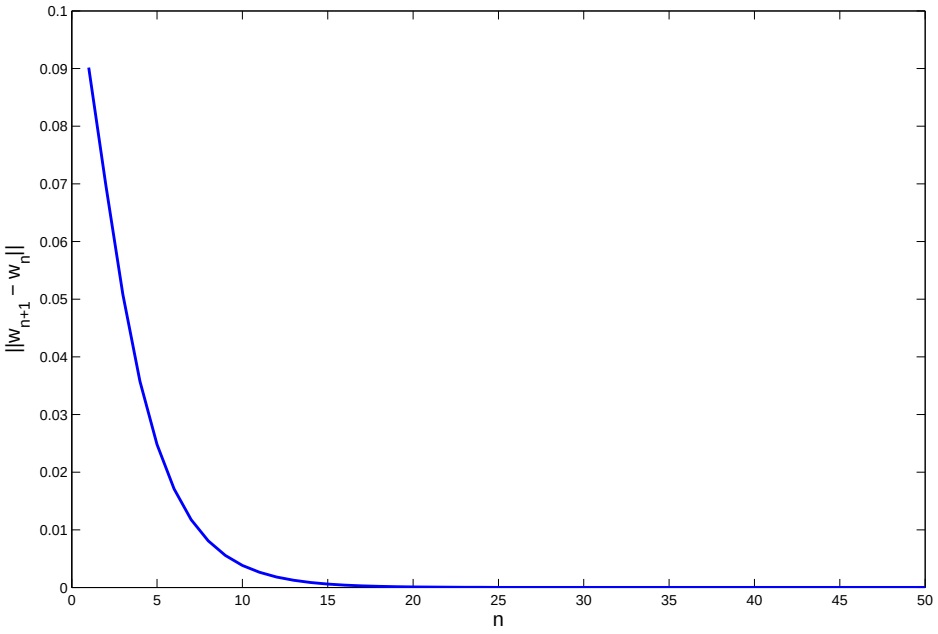

Figure 4.2: FastICA converges successfully.

this case, the empirical FastICA algorithm with respect to μ_N^1 and μ_N^2 performs better in terms of the convergence possibility; 2) If we consider only μ_N^3 and μ_N^4, then the kurtosis nonlinearity is the least reliable in terms of the convergence possibility. To achieve zero failure, "kurtosis" needs a sample size no less than 500, while "Gauss" and "tanh" require only $N = 100$. 3) If we have a very large sample size, say $N > 1000$, then the convergence is no longer a problem since the chance of convergence failure is less than 0.1%.

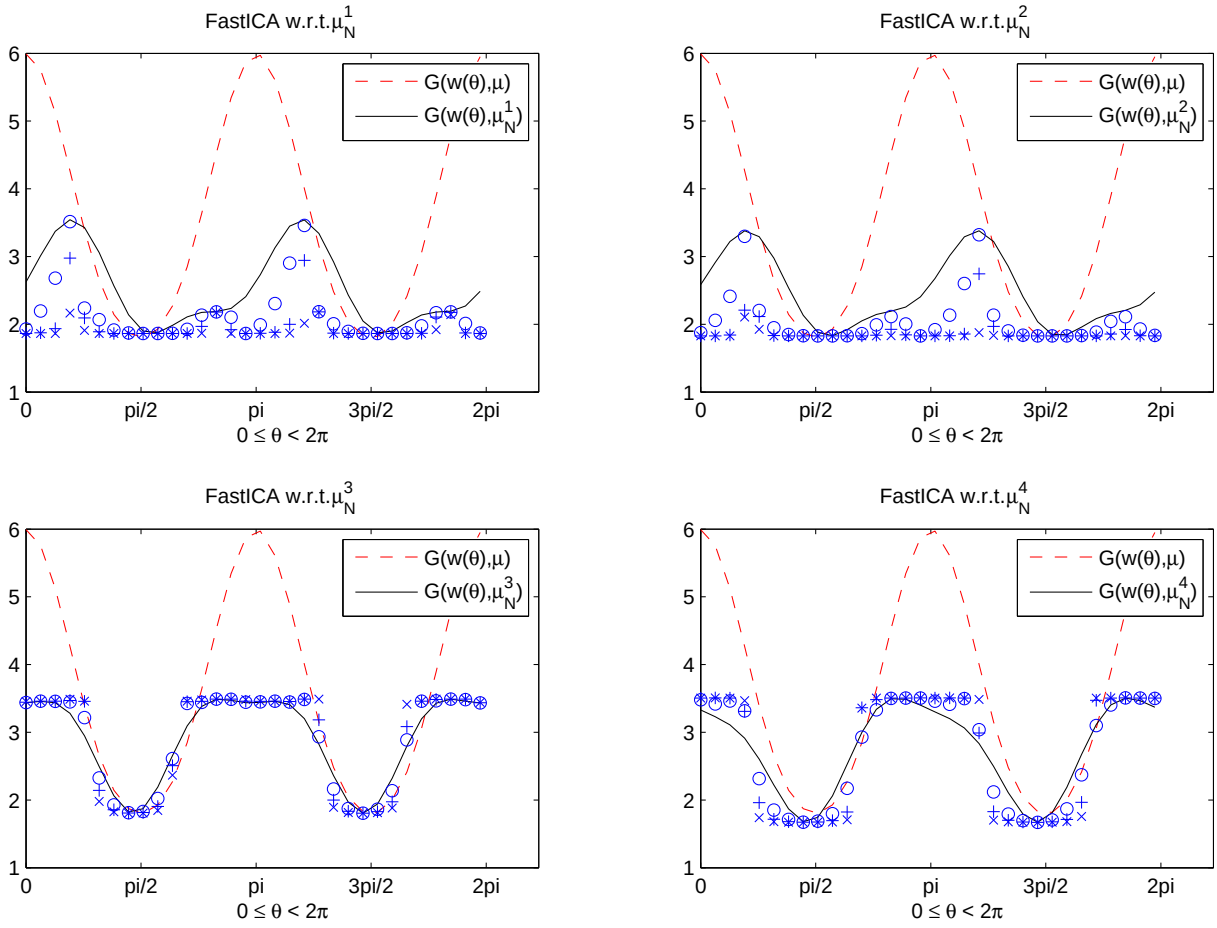

Figure 4.3: The empirical contrast function looks ill-conditioned due to small sample size ($N = 100$). However, empirical FastICA still converges as expected to a local minimizer of the respective contrast function.

Example 4.3.2. The purpose of this example is to numerically validate Theorem 4.2.4, which states that the FastICA algorithm yields a sequence that converges to a local minimizer of the respective contrast function. As what

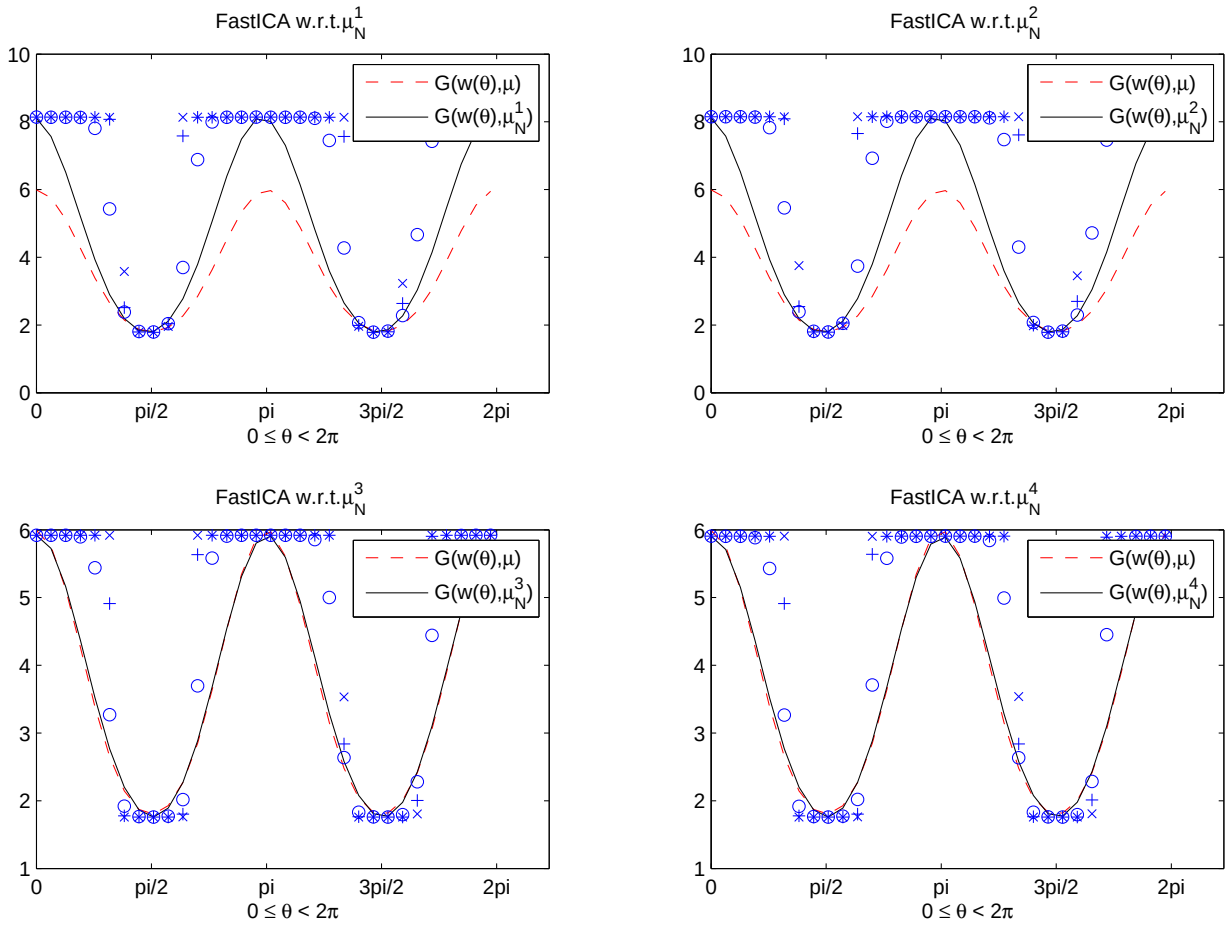

Figure 4.4: With moderate sample size ($N = 1000$), the empirical contrast function is relatively well-behaved. In this case, empirical FastICA gives good estimate.

we did in Example 3.4.2, we iterated FastICA three times for initial input $\mathbf{w}_0(\theta) = (\cos(\theta), \sin(\theta))$, with initial angle θ ranging from 0 to 2π. We recorded the outcome of each iteration, namely $\mathbf{w}_i(\theta) \stackrel{\text{def}}{=} \boldsymbol{f}(\mathbf{w}_{i-1}(\theta), \mu_N^k)$ for $i = 1, 2, 3$, and then plotted $G(\mathbf{w}_i(\theta), \mu_N^k)$, using marks "$\circ$", "$+$" and "$\times$", which correspond respectively to $i = 1, 2, 3$. In Fig 4.3, we used a small sample with $N = 100$, which leads to an ill-behaved $G(\cdot, \mu_N^k)$. However, from the graph we observe that the empirical FastICA algorithm could still converge to a local minimizer of $G(\cdot, \mu_N^k)$. In Fig 4.4, we increased the sample size to $N = 1000$. It is easily seen that, with increased sample size, $G(\cdot, \mu_N^k)$ is closer to $G(\cdot, \mu_N^k)$ and the empirical FastICA algorithm exhibits better performance, especially for $k = 3, 4$.

Example 4.3.3. In this example, we are interested in the convergence speed of the empirical FastICA algorithm. We refer to Example 3.4.4 for a comparison with the "theoretical" version of the algorithm. Here, we choose an arbitrary initial point $\mathbf{w}_0(\theta)$ near $\mathbf{e}_2$ and run FastICA for 9 iterations. In Fig 4.5, we plot $\|\mathbf{w}_{n+1} - \mathbf{a}_N^k\|/\|\mathbf{w}_n - \mathbf{a}_N^k\|$ for different sample size, namely, $N = 10^2, 10^3$ and 10^4. From the figure, we see that due to finite size of the sample, we only have a linear convergence speed since the ratio $\|\mathbf{w}_{n+1} - \mathbf{a}_N^k\|/\|\mathbf{w}_n - \mathbf{a}_N^k\|$ remains positive and stable during the entire simulation. Nevertheless, we also notice that the increase of the sample size N causes the drop of the ratio towards 0, which is logical for the reason that we would have a zero ratio if $N = +\infty$. Lastly, we point out that the magnitude of the ratio $\|\mathbf{w}_{n+1} - \mathbf{e}_2\|/\|\mathbf{w}_n - \mathbf{e}_2\|$ in Example 3.4.4 is at the level 10^{-4}, whereas here our $\|\mathbf{w}_{n+1} - \mathbf{a}_N^k\|/\|\mathbf{w}_n - \mathbf{a}_N^k\|$ for $N = 10^4$ is at around 10^{-2}.

4.4 Proof of Proposition 4.1.6

We will only provide the proof of the case $k = 1$ and $k = 4$, since the same approach applies easily to the case $k = 2, 3$. The proof will be divided into several parts. First, we will use the USLLN to prove (4.1.6)-(4.1.8) for the case $k = 1$, which is the keystone of the whole proof. Next, we will use these results to prove (4.1.6)-(4.1.8) for the case $k = 4$. Finally, we will show that (4.1.9) and (4.1.10) are direct consequences of (4.1.6)-(4.1.8).

4.4.1 Proof of (4.1.6)-(4.1.8) for $k = 1$.

We begin by proving (4.1.6)-(4.1.8) for the case $k = 1$. To show the uniform convergence, it suffices to verify that the functions $G(\mathbf{w}, \mu_N^1)$, $\boldsymbol{h}(\mathbf{w}, \mu_N^1)$ and $\nabla\boldsymbol{h}(\mathbf{w}, \mu_N^1)$ each satisfies the hypothesis of the USLLN. We claim that this

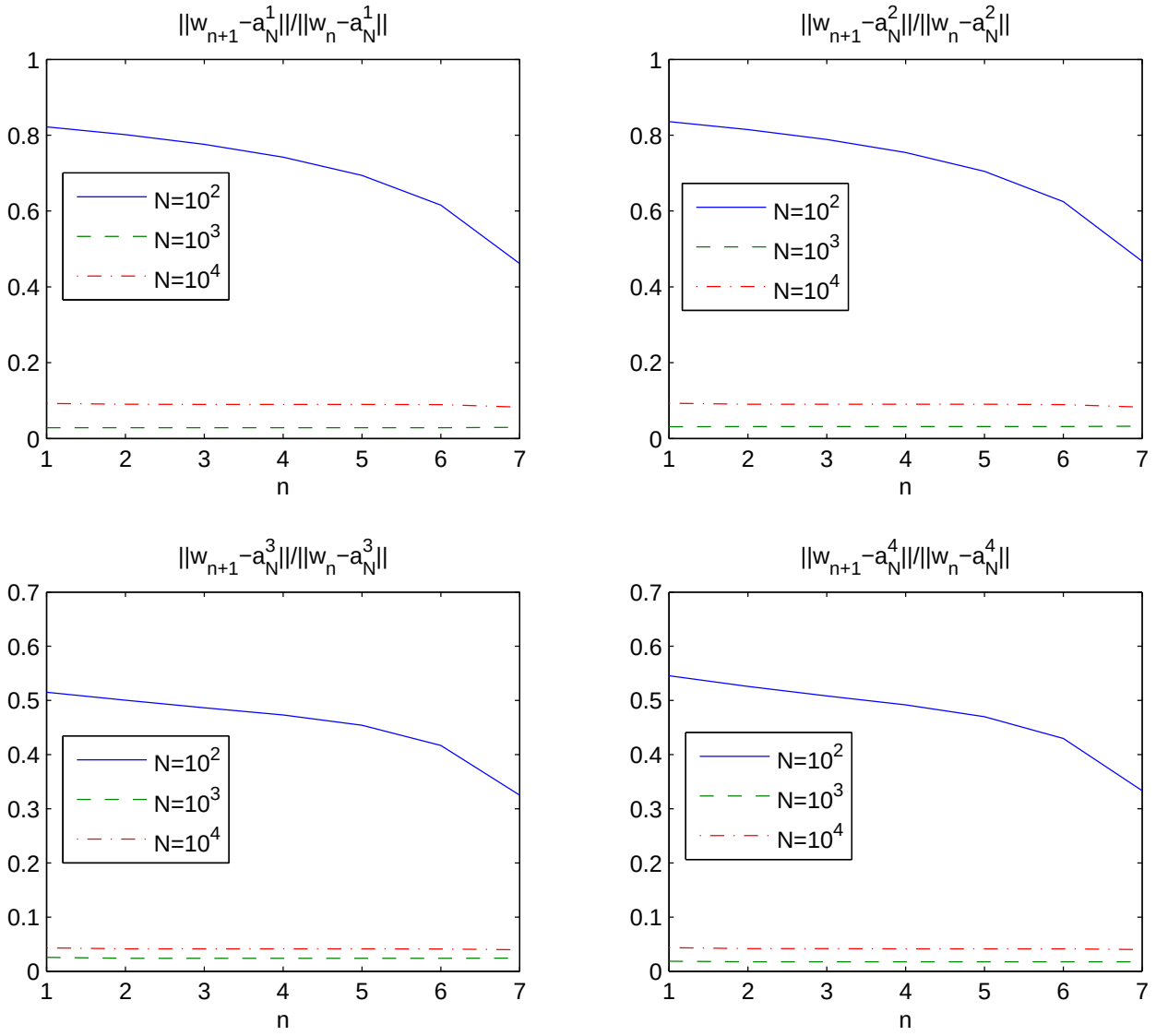

Figure 4.5: Convergence speed of the empirical FastICA algorithm.

is guaranteed by Assumption 3. In fact, we have

$$\begin{aligned}
\mathbb{E}\Big[\sup_{\mathbf{w}\in\mathcal{S}}|G(\mathbf{w}^{\mathsf{T}}\mathbf{x})|\Big] &\le \mathbb{E}\Big[\sup_{\mathbf{w}\in\mathcal{S}} c|\mathbf{w}^{\mathsf{T}}\mathbf{x}|^p\Big] \le c\mathbb{E}\Big[\|\mathbf{x}\|^p\Big] < \infty,\\
\mathbb{E}\Big[\sup_{\mathbf{w}\in\mathcal{S}}\|g'(\mathbf{w}^{\mathsf{T}}\mathbf{x})\mathbf{w} - g(\mathbf{w}^{\mathsf{T}}\mathbf{x})\mathbf{x}\|\Big] &\le \mathbb{E}\Big[\sup_{\mathbf{w}\in\mathcal{S}}\|g'(\mathbf{w}^{\mathsf{T}}\mathbf{x})\|\|\mathbf{w}\| + \|g(\mathbf{w}^{\mathsf{T}}\mathbf{x})\|\|\mathbf{x}\|\Big]\\
&\le \mathbb{E}\Big[\sup_{\mathbf{w}\in\mathcal{S}} c|\mathbf{w}^{\mathsf{T}}\mathbf{x}|^p\|\mathbf{w}\| + c|\mathbf{w}^{\mathsf{T}}\mathbf{x}|^p\|\mathbf{x}\|\Big]\\
&\le c\mathbb{E}\Big[\|\mathbf{x}\|^p + \|\mathbf{x}\|^{p+1}\Big] < \infty,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{\mathbf{w}\in\mathcal{S}}\|g''(\mathbf{w}^{\mathsf{T}}\mathbf{x})\mathbf{w}\mathbf{x}^{\mathsf{T}} + g'(\mathbf{w}^{\mathsf{T}}\mathbf{x})\mathbf{I} - g'(\mathbf{w}^{\mathsf{T}}\mathbf{x})\mathbf{x}\mathbf{x}^{\mathsf{T}}\|\Big]\\
\le\ &\mathbb{E}\Big[\sup_{\mathbf{w}\in\mathcal{S}}\|g''(\mathbf{w}^{\mathsf{T}}\mathbf{x})\|\|\mathbf{w}\|\|\mathbf{x}\| + \|g'(\mathbf{w}^{\mathsf{T}}\mathbf{x})\| + \|g'(\mathbf{w}^{\mathsf{T}}\mathbf{x})\|\|\mathbf{x}\|^2\Big]\\
\le\ &c\mathbb{E}\Big[\|\mathbf{x}\|^p + \|\mathbf{x}\|^{p+1} + \|\mathbf{x}\|^{p+2}\Big] < \infty.
\end{aligned}$$

Then the proof of (4.1.6)-(4.1.8) for the case $k = 1$ is achieved.

4.4.2 Proof of (4.1.6)-(4.1.8) for $k = 4$.

Note that we have

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{S}}\|G(\mathbf{w},\mu) - G(\mathbf{w},\mu_N^4)\|\\
\le\ &\sup_{\mathbf{w}\in\mathcal{S}}\Big(\|G(\mathbf{w},\mu) - G(\mathbf{w},\mu_N^1)\| + \|G(\mathbf{w},\mu_N^1) - G(\mathbf{w},\mu_N^4)\|\Big)\\
\le\ &\sup_{\mathbf{w}\in\mathcal{S}}\|G(\mathbf{w},\mu) - G(\mathbf{w},\mu_N^1)\| + \sup_{\mathbf{w}\in\mathcal{S}}\|G(\mathbf{w},\mu_N^1) - G(\mathbf{w},\mu_N^4)\|.
\end{aligned}$$

Hence, for (4.1.6) it suffices to show

$$\sup_{\mathbf{w}\in\mathcal{S}}\|G(\mathbf{w},\mu_N^1) - G(\mathbf{w},\mu_N^4)\| \xrightarrow[N\to\infty]{a.s.} 0. \tag{4.4.1}$$

Similarly, to prove (4.1.7) and (4.1.8) for the case $k = 4$, we have to show that

$$\sup_{\mathbf{w}\in\mathcal{S}}\|\boldsymbol{h}(\mathbf{w},\mu_N^1) - \boldsymbol{h}(\mathbf{w},\mu_N^4)\| \xrightarrow[N\to\infty]{a.s.} 0, \tag{4.4.2}$$

$$\sup_{\mathbf{w}\in\mathcal{S}}\|\nabla\boldsymbol{h}(\mathbf{w},\mu_N^1) - \nabla\boldsymbol{h}(\mathbf{w},\mu_N^4)\| \xrightarrow[N\to\infty]{a.s.} 0. \tag{4.4.3}$$

To proceed, we need the following result:

Lemma 4.4.1. *Let $\Phi(u,v,w)$ be a polynomial function of three variables u, v and w. Then we have the following convergence:*

$$\frac{1}{N}\sum_{t=1}^{N}\Phi(\|\mathbf{x}(t)\|, \|\mathbf{C}_N^{-1/2}\|, \|\bar{\mathbf{x}}\|) \xrightarrow[N\to\infty]{a.s.} \mathbb{E}\Big[\Phi\big(\|\mathbf{x}(1)\|, \|\mathbf{I}\|, 0\big)\Big]. \tag{4.4.4}$$

Consequently, for any polynomial function with positive coefficients $\widetilde{\Phi}$ *of two variables* u *and* v*, we have*

$$\frac{1}{N}\sum_{t=1}^{N}\widetilde{\Phi}\big(\|\mathbf{x}(t)\|,\|\mathbf{z}(t)\|\big)\|\mathbf{x}(t)-\mathbf{z}(t)\| \xrightarrow[N\to\infty]{a.s.} 0. \tag{4.4.5}$$

Proof of Lemma 4.4.1. From the SLLN, we have $\mathbf{C}_N^{-1/2} \xrightarrow[N\to\infty]{a.s.} \mathbf{I}$, $\bar{\mathbf{x}} \xrightarrow[N\to\infty]{a.s.} 0$, and $\frac{1}{N}\sum_{t=1}^{N}\|\mathbf{x}(t)\|^p \xrightarrow[N\to\infty]{a.s.} \mathbb{E}[\|\mathbf{x}(1)\|^p]<\infty$. Since $\mathbf{C}_N^{-1/2}$ and $\bar{\mathbf{x}}$ do not depend on t, we get (4.4.4). As for (4.4.5), we note that

$$\begin{aligned}
&\frac{1}{N}\sum_{t=1}^{N}\widetilde{\Phi}\big(\|\mathbf{x}(t)\|,\|\mathbf{z}(t)\|\big)\|\mathbf{x}(t)-\mathbf{z}(t)\| \\
\le&\frac{1}{N}\sum_{t=1}^{N}\widetilde{\Phi}\Big(\|\mathbf{x}(t)\|,\quad \|\mathbf{C}_N^{-\frac{1}{2}}\|\|\mathbf{x}(t)\|+\|\mathbf{C}_N^{-\frac{1}{2}}\|\|\bar{\mathbf{x}}\|\Big)\Big(\|\mathbf{C}_N^{-\frac{1}{2}}-\mathbf{I}\|\|\mathbf{x}(t)\|+\|\mathbf{C}_N^{-\frac{1}{2}}\|\|\bar{\mathbf{x}}\|\Big) \\
=&\left(\frac{1}{N}\sum_{t=1}^{N}\widetilde{\Phi}\Big(\|\mathbf{x}(t)\|,\quad \|\mathbf{C}_N^{-\frac{1}{2}}\|\|\mathbf{x}(t)\|+\|\mathbf{C}_N^{-\frac{1}{2}}\|\|\bar{\mathbf{x}}\|\Big)\|\mathbf{x}(t)\|\right)\|\mathbf{C}_N^{-\frac{1}{2}}-\mathbf{I}\| \\
&+\left(\frac{1}{N}\sum_{t=1}^{N}\widetilde{\Phi}\Big(\|\mathbf{x}(t)\|,\quad \|\mathbf{C}_N^{-\frac{1}{2}}\|\|\mathbf{x}(t)\|+\|\mathbf{C}_N^{-\frac{1}{2}}\|\|\bar{\mathbf{x}}\|\Big)\|\mathbf{C}_N^{-1/2}\|\right)\|\bar{\mathbf{x}}\|.
\end{aligned}$$

Then applying (4.4.4) and the fact $\|\mathbf{C}_N^{-\frac{1}{2}}-\mathbf{I}\| \xrightarrow[N\to\infty]{a.s.} 0$, $\|\bar{\mathbf{x}}\| \xrightarrow[N\to\infty]{a.s.} 0$, we achieve (4.4.5). □

Thanks to Lemma 4.4.1, now it suffices to prove that there exists a polynomial function Φ such that the terms on the left hand side of (4.4.1)-(4.4.3) can be bounded by $\frac{1}{N}\sum_{t=1}^{N}\Phi\big(\|\mathbf{x}(t)\|,\|\mathbf{z}(t)\|\big)\|\mathbf{x}(t)-\mathbf{z}(t)\|$. By the polynomial growth of $g(\cdot)$ (see Assumption 3), we deduce that

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{S}}\|G(\mathbf{w},\mu_N^1)-G(\mathbf{w},\mu_N^4)\| \\
=&\sup_{\mathbf{w}\in\mathcal{S}}\Big\|\frac{1}{N}\sum_{t=1}^{N}G(\mathbf{w}^\mathsf{T}\mathbf{x}(t))-\frac{1}{N}\sum_{i=1}^{N}G(\mathbf{w}^\mathsf{T}\mathbf{z}(t))\Big\| \\
=&\sup_{\mathbf{w}\in\mathcal{S}}\Big\|\frac{1}{N}\sum_{t=1}^{N}g\Big(\xi(t)\mathbf{w}^\mathsf{T}\mathbf{x}(t)+(1-\xi(t))\mathbf{w}^\mathsf{T}\mathbf{z}(t)\Big)\mathbf{w}^\mathsf{T}(\mathbf{x}(t)-\mathbf{z}(t))\Big\| \\
\le& c\sup_{\mathbf{w}\in\mathcal{S}}\Big(\frac{1}{N}\sum_{t=1}^{N}\Big|\xi(t)\mathbf{w}^\mathsf{T}\mathbf{x}(t)+(1-\xi(t))\mathbf{w}^\mathsf{T}\mathbf{z}(t)\Big|^p\|\mathbf{x}(t)-\mathbf{z}(t)\|\Big) \\
\le&\frac{c}{N}\sum_{t=1}^{N}\Big(\|\mathbf{x}(t)\|+\|\mathbf{z}(t)\|\Big)^p\|\mathbf{x}(t)-\mathbf{z}(t)\|,
\end{aligned}$$

where $\xi(t)$ is an intermediate value between 0 and 1. Similarly, we have

$$
\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{S}} \|\boldsymbol{h}(\mathbf{w},\mu_N^1) - \boldsymbol{h}(\mathbf{w},\mu_N^4)\| \\
&= \sup_{\mathbf{w}\in\mathcal{S}} \Big\| \frac{1}{N}\sum_{t=1}^{N} \Big(g'\big(\mathbf{w}^\mathsf{T}\mathbf{x}(t)\big)\mathbf{w} - g\big(\mathbf{w}^\mathsf{T}\mathbf{x}(t)\big)\mathbf{x}(t)\Big) \\
&\quad - \frac{1}{N}\sum_{t=1}^{N} \Big(g'\big(\mathbf{w}^\mathsf{T}\mathbf{z}(t)\big)\mathbf{w} - g\big(\mathbf{w}^\mathsf{T}\mathbf{z}(t)\big)\mathbf{z}(t)\Big)\Big\| \\
&= \sup_{\mathbf{w}\in\mathcal{S}} \Big\| \frac{1}{N}\sum_{t=1}^{N} \Big[\Big(g'\big(\mathbf{w}^\mathsf{T}\mathbf{x}(t)\big) - g'\big(\mathbf{w}^\mathsf{T}\mathbf{z}(t)\big)\Big)\mathbf{w} - \Big(g\big(\mathbf{w}^\mathsf{T}\mathbf{x}(t)\big)\mathbf{z}(t) - g\big(\mathbf{w}^\mathsf{T}\mathbf{z}(t)\big)\mathbf{z}(t)\Big) \\
&\quad - \Big(g\big(\mathbf{w}^\mathsf{T}\mathbf{x}(t)\big)\mathbf{x}(t) - g\big(\mathbf{w}^\mathsf{T}\mathbf{x}(t)\big)\mathbf{z}(t)\Big)\Big]\Big\| \\
&\le \sup_{\mathbf{w}\in\mathcal{S}} \Big[\Big\| \frac{1}{N}\sum_{t=1}^{N} g''\Big(\xi_1(t)\mathbf{w}^\mathsf{T}\mathbf{x}(t) + (1-\xi_1(t))\mathbf{w}^\mathsf{T}\mathbf{z}(t)\Big)\mathbf{w}^\mathsf{T}\big(\mathbf{x}(t)-\mathbf{z}(t)\big)\mathbf{w}\Big\| \\
&\quad + \Big\| \frac{1}{N}\sum_{t=1}^{N} g'\Big(\xi_2(t)\mathbf{w}^\mathsf{T}\mathbf{x}(t) + (1-\xi_2(t))\mathbf{w}^\mathsf{T}\mathbf{z}(t)\Big)\mathbf{w}^\mathsf{T}\big(\mathbf{x}(t)-\mathbf{z}(t)\big)\mathbf{z}(t)\Big\| \\
&\quad + \Big\| \frac{1}{N}\sum_{t=1}^{N} \Big(g(\mathbf{w}^\mathsf{T}\mathbf{x}(t))(\mathbf{x}(t)-\mathbf{z}(t))\Big)\Big\|\Big] \\
&\le \frac{1}{N}\sum_{t=1}^{N} \big(\|\mathbf{x}(t)\| + \|\mathbf{z}(t)\|\big)^p \big\|\mathbf{x}(t)-\mathbf{z}(t)\big\| + \frac{1}{N}\sum_{t=1}^{N} \big(\|\mathbf{x}(t)\| + \|\mathbf{z}(t)\|\big)^p \|\mathbf{z}(t)\| \big\|\mathbf{x}(t)-\mathbf{z}(t)\big\| \\
&\quad + \frac{1}{N}\sum_{t=1}^{N} \|\mathbf{x}(t)\|^p \big\|\mathbf{x}(t)-\mathbf{z}(t)\big\| \\
&\le \frac{1}{N}\sum_{t=1}^{N} \Big[\big(\|\mathbf{x}(t)\| + \|\mathbf{z}(t)\|\big)^p + \big(\|\mathbf{x}(t)\| + \|\mathbf{z}(t)\|\big)^p \|\mathbf{z}(t)\| + \|\mathbf{x}(t)\|^p\Big] \big\|\mathbf{x}(t)-\mathbf{z}(t)\big\|,
\end{aligned}
$$

where in the last line the term in the brackets is a polynomial of $\|\mathbf{x}(t)\|$ and $\|\mathbf{z}(t)\|$. Using the same approach again, we can show that

$$
\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{S}} \|\nabla\boldsymbol{h}(\mathbf{w},\mu_N^1) - \nabla\boldsymbol{h}(\mathbf{w},\mu_N^4)\| \\
&= \sup_{\mathbf{w}\in\mathcal{S}} \Bigg(\Big\| \frac{1}{N}\sum_{t=1}^{N} \Big[g''\big(\mathbf{w}^\mathsf{T}\mathbf{x}(t)\big)\mathbf{w}\mathbf{x}(t)^\mathsf{T} + g'\big(\mathbf{w}^\mathsf{T}\mathbf{x}(t)\big)\mathbf{I} - g'\big(\mathbf{w}^\mathsf{T}\mathbf{x}(t)\big)\mathbf{x}(t)\mathbf{x}(t)^\mathsf{T}\Big] \\
&\quad - \frac{1}{N}\sum_{t=1}^{N} \Big[g''\big(\mathbf{w}^\mathsf{T}\mathbf{z}(t)\big)\mathbf{w}\mathbf{z}(t)^\mathsf{T} + g'\big(\mathbf{w}^\mathsf{T}\mathbf{z}(t)\big)\mathbf{I} - g'\big(\mathbf{w}^\mathsf{T}\mathbf{z}(t)\big)\mathbf{x}(t)\mathbf{z}(t)^\mathsf{T}\Big]\Big\|\Bigg) \\
&\le \sup_{\mathbf{w}\in\mathcal{S}} \Bigg(\frac{1}{N}\sum_{t=1}^{N} \Big(\|g''\big(\mathbf{w}^\mathsf{T}\mathbf{x}(t)\big)\mathbf{w}\mathbf{x}(t)^\mathsf{T} - g''\big(\mathbf{w}^\mathsf{T}\mathbf{z}(t)\big)\mathbf{w}\mathbf{z}(t)^\mathsf{T}\| + \|g'\big(\mathbf{w}^\mathsf{T}\mathbf{x}(t)\big)\mathbf{I} - g'\big(\mathbf{w}^\mathsf{T}\mathbf{z}(t)\big)\mathbf{I}\|
\end{aligned}
$$

$$+\left\|g'\big(\mathbf{w}^{\mathsf{T}}\mathbf{x}(t)\big)\mathbf{x}(t)\mathbf{x}(t)^{\mathsf{T}} - g'\big(\mathbf{w}^{\mathsf{T}}\mathbf{z}(t)\big)\mathbf{z}(t)\mathbf{z}(t)^{\mathsf{T}}\right\|\Big)\Big)$$

$$\leq \frac{1}{N}\sum_{t=1}^{N}\Big(\Phi_1(\|\mathbf{x}\|,\|\mathbf{z}(t)\|)+\Phi_2(\|\mathbf{x}\|,\|\mathbf{z}(t)\|)+\Phi_3(\|\mathbf{x}\|,\|\mathbf{z}(t)\|)\Big)\|\mathbf{x}(t)-\mathbf{z}(t)\|,$$

where Φ_i are polynomial functions, $i = 1, 2, 3$.

4.4.3 Proof of (4.1.9) and (4.1.10)

Let us first introduce the following lemma:

Lemma 4.4.2. *Let $\{\mathbf{y}_N(\cdot)\}$ be a sequence of functions $\mathbb{R}^d \mapsto \mathbb{R}^m$, such that*

$$\sup_{\theta\in\Theta}\|\mathbf{y}_N(\theta)-\mathbf{y}(\theta)\| \xrightarrow[N\to\infty]{} 0, \tag{4.4.6}$$

where Θ is a compact subset of $\mathbb{R}^d$ and $\mathbf{y}(\cdot)$ is a continuous function. Let $\mathcal{P}:\mathbb{R}^m\to\mathbb{R}^n$ be a derivable mapping such that $\nabla\mathcal{P}(\cdot)$ is continuous in the set $\mathcal{V}=\{\mathbf{c}:\inf_{\theta\in\Theta}\|\mathbf{c}-\mathbf{y}(\theta)\|\leq\epsilon\}$ for some $\epsilon>0$. Then we have

$$\sup_{\theta\in\Theta}\|\mathcal{P}(\mathbf{y}_N(\theta))-\mathcal{P}(\mathbf{y}(\theta))\| \xrightarrow[N\to\infty]{} 0. \tag{4.4.7}$$

Proof. Let us denote the ith component of $\mathcal{P}$ by $\mathcal{P}_i$. It suffices to prove that

$$\sup_{\theta\in\Theta}|\mathcal{P}_i(\mathbf{y}_N(\theta))-\mathcal{P}_i(\mathbf{y}(\theta))| \xrightarrow[N\to\infty]{} 0, \quad i=1,\ldots,n. \tag{4.4.8}$$

We have

$$\begin{aligned}
&\sup_{\theta\in\Theta}|\mathcal{P}_i(\mathbf{y}_N(\theta))-\mathcal{P}_i(\mathbf{y}(\theta))|\\
=&\sup_{\theta\in\Theta}\Big|\nabla\mathcal{P}_i\Big(c\mathbf{y}_N(\theta)+(1-c)\mathbf{y}(\theta)\Big)\Big(\mathbf{y}_N(\theta)-\mathbf{y}(\theta)\Big)\Big|\\
\leq&\sup_{\theta\in\Theta}\Big\|\nabla\mathcal{P}_i\Big(c\mathbf{y}_N(\theta)+(1-c)\mathbf{y}(\theta)\Big)\Big\|\times\sup_{\theta\in\Theta}\Big\|\mathbf{y}_N(\theta)-\mathbf{y}(\theta)\Big\|.
\end{aligned}$$

Denote $\mathbf{u}_N(\theta)\overset{\text{def}}{=}c\mathbf{y}_N(\theta)+(1-c)\mathbf{y}(\theta)$. Then by (4.4.6), we have $\sup_{\theta\in\Theta}\|\mathbf{u}_N(\theta)-\mathbf{y}(\theta)\| \xrightarrow[N\to\infty]{} 0$. Then it follows from the continuity of $\nabla\mathcal{P}$ and the compactness of $\mathcal{V}$ that

$$\sup_{\theta\in\Theta}\big\|\nabla\mathcal{P}_i\big(\mathbf{u}_N(\theta)\big)\big\|\leq\sup_{\mathbf{u}\in\mathcal{V}}\big\|\nabla\mathcal{P}_i(\mathbf{u})\big\|<\infty$$

holds almost surely for large N. Thus we achieved (4.4.8), and (4.4.7) follows. □

To prove (4.1.9), let us consider the mapping $\mathcal{P}^1 : \mathbb{R}^d \to \mathbb{R}^d$ defined by $\mathcal{P}^1(\mathbf{x}) \stackrel{\text{def}}{=} \mathbf{x}/\|\mathbf{x}\|$. Clearly, $\mathcal{P}^1$ has continuous derivative in $\mathbb{R}^d/\{\mathbf{0}\}$. Besides, we have $\boldsymbol{f}(\mathbf{w},\nu) = \mathcal{P}^1(\boldsymbol{h}(\mathbf{w},\nu))$ for any measure ν. Note that $\|\boldsymbol{h}(\mathbf{w},\mu)\| \neq 0$ in $\mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}$ since $\boldsymbol{h}(\cdot,\mu)$ is continuous and $\|\boldsymbol{h}(\mathbf{a},\mu)\| = |H(\mathbf{a},\mu)| > 0$. Then by Lemma 4.4.2 and (4.1.7), we get

$$\begin{aligned} &\sup_{\mathbf{w}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}} \|\boldsymbol{f}(\mathbf{w},\mu_N^k) - \boldsymbol{f}(\mathbf{w},\mu)\| \\ =\ &\sup_{\mathbf{w}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}} \|\mathcal{P}^1\big(\boldsymbol{h}(\mathbf{w},\mu_N^k)\big) - \mathcal{P}^1\big(\boldsymbol{h}(\mathbf{w},\mu)\big)\| \xrightarrow[N\to\infty]{a.s.} 0. \end{aligned}$$

As for (4.1.10), we consider the mapping $\mathcal{P}^2 : \mathbb{R}^d \times \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ defined by

$$\mathcal{P}^2(\mathbf{x},\mathbf{y}) \stackrel{\text{def}}{=} \frac{(\|\mathbf{x}\|^2\mathbf{I} - \mathbf{x}\mathbf{x}^\mathsf{T})\mathbf{y}}{\|\mathbf{x}\|^3}.$$

Clearly, $\mathcal{P}^2$ is continuous in $\mathbb{R}^d \times \mathbb{R}^{d\times d}/\{\mathbf{0}\times\mathbb{R}^{d\times d}\}$ and it can be considered as a mapping from $\mathbb{R}^{d^2+d}$ to $\mathbb{R}^{d^2}$. Applying Lemma 4.4.2, we get

$$\begin{aligned} &\sup_{\mathbf{w}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}} \|\nabla\boldsymbol{f}(\mathbf{w},\mu_N^k) - \nabla\boldsymbol{f}(\mathbf{w},\mu)\| \\ =\ &\sup_{\mathbf{w}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}} \Big\|\mathcal{P}^2\Big(\boldsymbol{h}(\mathbf{w},\mu_N^k), \nabla\boldsymbol{h}(\mathbf{w},\mu_N^k)\Big) - \mathcal{P}^2\Big(\boldsymbol{h}(\mathbf{w},\mu), \nabla\boldsymbol{h}(\mathbf{w},\mu)\Big)\Big\| \xrightarrow[N\to\infty]{a.s.} 0. \end{aligned}$$

Chapter 5

Asymptotic Analysis of FastICA estimators

Contents

In Chapter 2 we showed that the column $\mathbf{a}$ of the mixing matrix $\mathbf{A}$ is a local minimizer of the theoretical contrast function $G(\mathbf{w}, \mu) \stackrel{\text{def}}{=} \mathbb{E}_{\mu}[G(\mathbf{w}^{\mathsf{T}}\mathbf{x})]$ on $\mathcal{S}$. In Chapter 3 we proposed four estimators $G(\mathbf{w}, \mu_N^k) \stackrel{\text{def}}{=} \mathbb{E}_{\mu_N^k}[G(\mathbf{w}^{\mathsf{T}}\boldsymbol{z})]$ of the contrast function $G(\mathbf{w}, \mu)$, and showed that for each k, the empirical contrast function $G(\mathbf{w}, \mu_N^k)$ has a local minimizer $\mathbf{a}_N^k$ near the column $\mathbf{a}$ on $\mathcal{S}$. We also proved the consistency of the empirical FastICA estimator by showing $\mathbf{a}_N^k \to \mathbf{a}$ almost surely as N tends to infinity. These facts suggest that the empirical FastICA estimator is actually an M-estimator. The aim of this chapter is to use the theory of M-estimator to derive the asymptotic normality and the asymptotic covariance matrix of our four empirical FastICA estimators.

5.1 Statement of the main result

We start by announcing the main result of this chapter. The following notations are adopted to simplify the formula:

$$\begin{aligned}
\eta_i &\stackrel{\text{def}}{=} \mathbb{E}[g(s_i)] \\
\alpha_i &\stackrel{\text{def}}{=} \mathbb{E}[g(s_i)s_i] \\
\rho_i &\stackrel{\text{def}}{=} \mathbb{E}[g'(s_i)] \\
\beta_i &\stackrel{\text{def}}{=} \mathbb{E}[g(s_i)^2]
\end{aligned}$$

$$\kappa_i \overset{\text{def}}{=} \frac{1}{4}(\mathbb{E}[s_i^4] - 1).$$

Besides, we recall that

$$\mathbf{Q}_N \overset{\text{def}}{=} \frac{1}{N}\sum_{t=1}^{N} \mathbf{x}(t)\mathbf{x}(t)^\mathsf{T} \quad \mathbf{C}_N \overset{\text{def}}{=} \mathbf{Q}_N - \bar{\mathbf{x}}\bar{\mathbf{x}}^\mathsf{T}.$$

Theorem 5.1.1. *Let $\mathbf{a}_i$ be the i-th column of the mixing matrix $\mathbf{A}$. We have*

$$N^{1/2}(\mathbf{a}_N^1 - \mathbf{a}_i) \xrightarrow[N\to\infty]{\mathscr{D}} \boldsymbol{z}_1 \tag{5.1.1}$$

$$N^{1/2}(\mathbf{a}_N^2 - \mathbf{a}_i) \xrightarrow[N\to\infty]{\mathscr{D}} \boldsymbol{z}_2 \tag{5.1.2}$$

$$N^{1/2}(\mathbf{Q}_N^{-1/2}\mathbf{a}_N^3 - \mathbf{a}_i) \xrightarrow[N\to\infty]{\mathscr{D}} \boldsymbol{z}_3 \tag{5.1.3}$$

$$N^{1/2}(\mathbf{C}_N^{-1/2}\mathbf{a}_N^4 - \mathbf{a}_i) \xrightarrow[N\to\infty]{\mathscr{D}} \boldsymbol{z}_4, \tag{5.1.4}$$

where $\boldsymbol{z}_k$ is a Gaussian random vector for each $k = 1, 2, 3, 4$. Moreover, we have

$$\mathrm{Cov}(\boldsymbol{z}_1) = \frac{\beta_i}{(\rho_i - \alpha_i)^2}(\mathbf{I} - \mathbf{a}_i\mathbf{a}_i^\mathsf{T}) \tag{5.1.5}$$

$$\mathrm{Cov}(\boldsymbol{z}_2) = \frac{\beta_i + 3\eta_i^2}{(\rho_i - \alpha_i)^2}(\mathbf{I} - \mathbf{a}_i\mathbf{a}_i^\mathsf{T}) \tag{5.1.6}$$

$$\mathrm{Cov}(\boldsymbol{z}_3) = \frac{\beta_i - \alpha_i^2}{(\rho_i - \alpha_i)^2}\big(\mathbf{I} - \mathbf{a}_i\mathbf{a}_i^\mathsf{T}\big) + \kappa_i\mathbf{a}_i\mathbf{a}_i^\mathsf{T} \tag{5.1.7}$$

$$\mathrm{Cov}(\boldsymbol{z}_4) = \frac{\beta_i - \alpha_i^2 + 3\eta_i^2}{(\rho_i - \alpha_i)^2}\big(\mathbf{I} - \mathbf{a}_i\mathbf{a}_i^\mathsf{T}\big) + \kappa_i\mathbf{a}_i\mathbf{a}_i^\mathsf{T}. \tag{5.1.8}$$

Proof. The proof of the theorem will be given in Section 5.5.2. It relies on the method of Lagrange multipliers and the theory of M-estimators, which will be introduced in the following sections. □

Remark 5.1.2. We notice that if s_i has symmetric distribution and if the nonlinearity $G(\cdot)$ is an even function, then the quantity η_i vanishes. In this case, we have $\mathrm{Cov}(\boldsymbol{z}_1) = \mathrm{Cov}(\boldsymbol{z}_2)$ and $\mathrm{Cov}(\boldsymbol{z}_3) = \mathrm{Cov}(\boldsymbol{z}_4)$.

The asymptotic normality (5.1.1) and (5.1.2) involve directly the FastICA estimator $\mathbf{a}_N^1$ and $\mathbf{a}_N^2$, while (5.1.3) and (5.1.4) concern $\mathbf{Q}_N^{-1/2}\mathbf{a}_N^3$ and $\mathbf{C}_N^{-1/2}\mathbf{a}_N^4$. This phenomenon can be justified as follows. If the mixing matrix $\mathbf{A}$ is not orthogonal, then we need to whiten the observed signal before implementing FastICA. By whitening the data, we transform the model into $\tilde{\mathbf{x}}(t) = \widetilde{\mathbf{A}}\tilde{s}(t)$, where $\widetilde{\mathbf{A}} = \mathbf{C}_N^{-1/2}\mathbf{A}$, as is explained in Section 2.1.2. The empirical FastICA algorithm actually gives an estimate of a row $(\mathbf{a}_N^4)^\mathsf{T}$

of $\widetilde{\mathbf{A}}^{-1} = \mathbf{A}^{-1}\mathbf{C}_N^{1/2}$. Hence, to find the corresponding row of the original demixing matrix $\mathbf{A}^{-1}$, one needs to take $(\mathbf{a}_N^4)^\mathsf{T}\mathbf{C}_N^{-1/2}$, or $\mathbf{C}_N^{-1/2}\mathbf{a}_N^4$.

The inconvenience of (5.1.5)-(5.1.8) is that they depend on the mixing matrix $\mathbf{A}$. If we are to propose a performance criterion for the choice of nonlinearity, it is better to use the *gain matrix*.

Definition 5.1.3. *Let $\widehat{\mathbf{W}}$ be an estimate of $\mathbf{A}^{-1}$. The gain matrix $\widehat{\mathbf{G}}$ associated to $\widehat{\mathbf{W}}$ is defined as $\widehat{\mathbf{G}} \overset{\text{def}}{=} \widehat{\mathbf{W}}\mathbf{A}$.*

Let $\mathbf{G}^{(k)}$ be the gain matrix of the FastICA estimator with respect to measure μ_N^k for $k = 1, 2, 3, 4$ and its (i, j)-entry be denoted by $\mathbf{G}_{ij}^{(k)}$. By Definition 5.1.3, we have $\mathbf{G}_{ij}^{(k)} = (\mathbf{a}_N^k)^\mathsf{T}\mathbf{a}_j$ for $k = 1, 2$; $\mathbf{G}_{ij}^{(3)} = (\mathbf{a}_N^3)^\mathsf{T}\mathbf{Q}_N^{-1/2}\mathbf{a}_j$ for $k = 3$ and $\mathbf{G}_{ij}^{(4)} = (\mathbf{a}_N^4)^\mathsf{T}\mathbf{C}_N^{-1/2}\mathbf{a}_j$ for $k = 4$, where $\mathbf{a}_N^k$ is the FastICA estimator of $\mathbf{a}_i$. From Theorem 5.1.1, we get immediately the following corollary:

Corollary 5.1.4. *Let $\mathbf{G}^{(k)}$ be the gain matrix of the FastICA estimator with respect to measure μ_N^k for $k = 1, 2, 3, 4$. We denote by $\mathbf{G}_{ij}^{(k)}$ the (i, j)-entry of $\mathbf{G}^{(k)}$. For $i = 1, \ldots, d$ and $j \neq i$ we have*

$$N^{1/2}\mathbf{G}_{ij}^{(k)} \xrightarrow[N\to\infty]{\mathscr{D}} \mathcal{N}(0, V_{ij}^{(k)})$$

where V_{ij}^k is given by

$$V_{ij}^{(k)} = \frac{\beta_i}{(\rho_i - \alpha_i)^2} \tag{5.1.9}$$

$$V_{ij}^{(k)} = \frac{\beta_i + 3\eta_i^2}{(\rho_i - \alpha_i)^2} \tag{5.1.10}$$

$$V_{ij}^{(k)} = \frac{\beta_i - \alpha_i^2}{(\rho_i - \alpha_i)^2} \tag{5.1.11}$$

$$V_{ij}^{(k)} = \frac{\beta_i - \alpha_i^2 + 3\eta_i^2}{(\rho_i - \alpha_i)^2}. \tag{5.1.12}$$

5.1.1 Related works

Several studies (Hyvärinen, 1997; Tichavsky et al., 2006; Shimizu, Hyvärinen, Yutaka, Hoyer, & Kerminen, 2006; Ollila, 2010; Reyhani, Ylipaavalniemi, Vigario, & Oja, 2012) concerning the asymptotic behavior of FastICA already exist.

The first attempt to calculate the asymptotic covariance matrix of FastICA estimator was due to Hyvärinen (Hyvärinen, 1997). In his paper, he derived the trace of the asymptotic covariance matrix of his FastICA estimator:

$$C(\mathbf{A})\frac{\beta_i - \alpha_i^2}{(\rho_i - \alpha_i)^2},$$

where $C(\mathbf{A})$ is a constant that depends only on the mixing matrix $\mathbf{A}$. The drawback of this work is that the author's whole argument lacks mathematical rigor.

Tichavsky, Koldovsky and Oja also tackled this subject in (Tichavsky et al., 2006). In their work, they studied the asymptotic property of the gain matrix of FastICA, for both one-unit and symmetrical version, and compared it with the theoretical Cramer-Rao bound. To derive their result which identical to (5.1.11), the authors proposed to study the output of FastICA after exactly one iteration with an ideal initialization $\mathbf{w}_0 = \mathbf{a}_i$. It was argued in the paper that although this one-iteration output is not really the FastICA estimator, it still somehow reflects latter's precision. Following this idea, the author derived the asymptotic variance of the gain matrix while taking into account of the effect of the data centering and whitening (equivalent to use measure μ_N^4). Their formula is correct for the case that the nonlinearity is an odd-function and the source signals have symmetric distribution, but in our opinion, it cannot be justified that their approach really yields the true asymptotic variance of the gain matrix.

Another related paper is (Ollila, 2010). Using a heuristic method based on the influence function (IF), the author derived a compact closed-form expression of the asymptotic covariance matrix of the general l-unit empirical FastICA estimator (i.e. the lth sequentially obtained FastICA estimator. By the uncorrelation principle, it must be orthogonal to all the previously obtained FastICA estimators):

$$\mathbf{\Sigma}_i = \sum_{j=1}^{i-1} \Big(\frac{\beta_j - \alpha_j^2}{(\rho_j - \alpha_j)^2} + 1 \Big) \mathbf{a}_j \mathbf{a}_j^\mathsf{T} + \kappa_i \mathbf{a}_i \mathbf{a}_i^\mathsf{T} + \frac{\beta_i - \alpha_i^2}{(\rho_i - \alpha_i)^2} \sum_{j=i+1}^{d} \mathbf{a}_j \mathbf{a}_j^\mathsf{T},$$

where it is assumed that the source signals are recovered in the order of $s_1, \ldots, s_d$. In the special case of $l = 1$, which is the case of one-unit FastICA, this formula is reduced to

$$\begin{aligned} \mathbf{\Sigma}_1 &= \kappa_1 \mathbf{a}_1 \mathbf{a}_1^\mathsf{T} + \frac{\beta_1 - \alpha_1^2}{(\rho_1 - \alpha_1)^2} \sum_{j=2}^{d} \mathbf{a}_j \mathbf{a}_j^\mathsf{T} \\ &= \kappa_1 \mathbf{a}_1 \mathbf{a}_1^\mathsf{T} + \frac{\beta_1 - \alpha_1^2}{(\rho_1 - \alpha_1)^2} (\mathbf{I} - \mathbf{a}_1 \mathbf{a}_1^\mathsf{T}), \end{aligned}$$

which coincides with our result (5.1.7). The drawback of this result is that it stems from a heuristic argument and it did not take into account the centering procedure neither.

The latest attempt to this subject is (Reyhani et al., 2012). The authors did not consider data centering or whitening (that is, their setting is equivalent to μ_N^1 in our case), made some assumptions that are unnecessarily strong, such as bounded sources, and obtained some results that seem erroneous. We cite their work here because the authors used the same approach

of M-estimator as is employed in this thesis to establish the asymptotic normality of the FastICA estimator.

5.2 Method of Lagrange multipliers

In mathematical optimization, the method of Lagrange multipliers provides a strategy for finding the local minima of a function subject to an equality constraints. For a detailed discussion about this subject, we refer to (Luenberger & Ye, 2008; Nesterov, 2004). Consider the optimization problem

$$\begin{aligned} &\text{minimize } \Phi(\mathbf{w}) \\ &\text{subject to } \mathbf{q}(\mathbf{w}) = 0, \end{aligned}$$

where $\Phi(\mathbf{w})$ is a mapping from $\mathbb{R}^n$ to $\mathbb{R}$ and $\mathbf{q}(\mathbf{w})$ is a mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$. We are going to recall the necessary condition for a point to be local minimizer of $\Phi(\cdot)$ subject to equality constraint $\mathbf{q}(\cdot)$. Before stating the result, we define the Lagrange function associated to the optimization problem:

$$\mathscr{L}(\mathbf{w}, \boldsymbol{\lambda}) = \Phi(\mathbf{w}) + \boldsymbol{\lambda}^{\mathsf{T}} \mathbf{q}(\mathbf{w}). \tag{5.2.1}$$

The following theorem is well-known.

Theorem 5.2.1 (First-Order Necessary Conditions)**.** *Let $\mathbf{w}^*$ be a local minimizer of $\Phi(\cdot)$ subject to the constraint $\mathbf{q}(\mathbf{w}) = 0$. Assume further that $\mathbf{w}^*$ is a regular point, i.e.*

$$\nabla \mathbf{q}(\mathbf{w}^*) \neq 0.$$

Then there exists $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ called Lagrange multiplier such that $(\mathbf{w}^*, \boldsymbol{\lambda}^*)$ is a stationary point of the Lagrange function (5.2.1). More precisely, the couple $(\mathbf{w}^*, \boldsymbol{\lambda}^*)$ satisfies

$$\begin{aligned} \partial_{\mathbf{w}} \mathscr{L}(\mathbf{w}^*, \boldsymbol{\lambda}^*) &= 0 \\ \partial_{\boldsymbol{\lambda}} \mathscr{L}(\mathbf{w}^*, \boldsymbol{\lambda}^*) &= 0. \end{aligned}$$

Let us apply this theorem to the optimization of our contrast functions:

$$\mathbf{a} = \underset{\mathbf{w} \in \mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}}{\operatorname{argmin}} \; G(\mathbf{w}, \mu), \tag{5.2.2}$$

$$\mathbf{a}_N^k = \underset{\mathbf{w} \in \mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}}{\operatorname{argmin}} \; G(\mathbf{w}, \mu_N^k). \tag{5.2.3}$$

5.2.1 Lagrange function of optimization problem (5.2.2)

Here, $\Phi(\mathbf{w}) = G(\mathbf{w}, \mu)$ and $\mathbf{q}(\mathbf{w}) = \|\mathbf{w}\|^2 - 1$. The Lagrange function is given by

$$\mathscr{L}(\mathbf{w}, \lambda) = G(\mathbf{w}, \mu) + \lambda(\|\mathbf{w}\|^2 - 1).$$

By Theorem 5.2.1, we get the following first order condition:

Lemma 5.2.2. *There exists $\lambda_{\mathbf{a}}$ such that $(\mathbf{a}, \lambda_{\mathbf{a}})$ is a solution of the equation $\mathbb{E}[\boldsymbol{\psi}^1(\mathbf{x}, \boldsymbol{\theta})] = 0$, where $\boldsymbol{\theta} \stackrel{\text{def}}{=} (\mathbf{w}, \lambda)$ and*

$$\boldsymbol{\psi}^1(\mathbf{x}, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \begin{bmatrix} g\big(\mathbf{w}^\mathsf{T}\mathbf{x}\big)\mathbf{x} + 2\lambda\mathbf{w} \\ \|\mathbf{w}\|^2 - 1 \end{bmatrix}. \tag{5.2.4}$$

Proof. We recall that $G(\mathbf{w}, \mu) = \mathbb{E}[G(\mathbf{w}^\mathsf{T}\mathbf{x})]$. Then we have the first order condition

$$\begin{aligned} \mathbb{E}[g(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{x}] + 2\lambda_{\mathbf{a}}\mathbf{a} &= 0 \\ \|\mathbf{a}\|^2 - 1 &= 0. \end{aligned} \tag{5.2.5}$$

Multiplying $\mathbf{a}^\mathsf{T}$ from left in (5.2.5), we obtain

$$\lambda_{\mathbf{a}} = -\frac{1}{2}\mathbb{E}[g\big(\mathbf{a}^\mathsf{T}\mathbf{x}\big)\mathbf{a}^\mathsf{T}\mathbf{x}].$$

□

5.2.2 Lagrange function of optimization problem (5.2.3)

Lagrange function for the case $k = 1$.
This case is similar to (5.2.2). We have

$$\mathbf{a}_N^1 = \operatorname*{argmin}_{\mathbf{w} \in \mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}} G(\mathbf{w}, \mu_N^1) = \operatorname*{argmin}_{\mathbf{w} \in \mathscr{B}_r(\mathbf{a}) \cap \mathcal{S}} \frac{1}{N}\sum_{t=1}^{N} G(\mathbf{w}^\mathsf{T}\mathbf{x}(t)).$$

Then we can deduce immediately the Lagrange function:

$$\mathscr{L}_1(\mathbf{w}, \lambda) = \frac{1}{N}\sum_{t=1}^{N} G(\mathbf{w}^\mathsf{T}\mathbf{x}(t)) + \lambda(\|\mathbf{w}\|^2 - 1).$$

Using the same proof as that of Lemma 5.2.2, we get the following result.

Lemma 5.2.3. *There exists λ_N^1 such that $\boldsymbol{\theta}_N^1 = (\mathbf{a}_N^1, \lambda_N^1)$ is a solution of the equation $\dfrac{1}{N}\displaystyle\sum_{t=1}^{N} \boldsymbol{\psi}^1(\mathbf{x}, \boldsymbol{\theta}) = 0$, where $\boldsymbol{\psi}^1(\mathbf{x}, \boldsymbol{\theta})$ is defined in (5.2.4) and*

$$\lambda_N^1 = -\frac{1}{2N}\sum_{t=1}^{N} g\Big((\mathbf{a}_N^1)^\mathsf{T}\mathbf{x}(t)\Big)(\mathbf{a}_N^1)^\mathsf{T}\mathbf{x}(t).$$

Lagrange function for the case $k = 2$.
We have

$$\mathbf{a}_N^2 = \underset{\mathbf{w}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}}{\operatorname{argmin}} G(\mathbf{w}, \mu_N^2) = \underset{\mathbf{w}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}}{\operatorname{argmin}} \frac{1}{N}\sum_{t=1}^{N} G(\mathbf{w}^\mathsf{T}(\mathbf{x}(t) - \bar{\mathbf{x}})). \quad (5.2.6)$$

Here, $\Phi(\mathbf{w}) = G(\mathbf{w}, \mu_N^2)$ and $\mathbf{q}(\mathbf{w}) = \|\mathbf{w}\|^2 - 1$. However, we prefer to introduce the auxiliary constraint $\bar{\mathbf{x}} = \mathbf{m}$, and write the optimization problem in the following equivalent form:

$$\begin{aligned} &\text{minimize} \quad \Phi(\mathbf{w}, \mathbf{m}) = \frac{1}{N}\sum_{t=1}^{N} G\big(\mathbf{w}^\mathsf{T}(\mathbf{x}(t) - \mathbf{m})\big) \\ &\text{subject to} \quad \mathbf{q}(\mathbf{w}, \mathbf{m}) = \begin{bmatrix} \|\mathbf{w}\|^2 - 1 \\ \bar{\mathbf{x}} - \mathbf{m} \end{bmatrix} = \mathbf{0}. \end{aligned} \quad (5.2.7)$$

The reason for this will be explained in the next section. Now, the corresponding Lagrange function is given by

$$\mathscr{L}_2(\mathbf{w}, \mathbf{m}, \lambda, \mathbf{k}) \overset{\text{def}}{=} \frac{1}{N}\sum_{t=1}^{N} G(\mathbf{w}^\mathsf{T}(\mathbf{x}(t) - \mathbf{m})) + \lambda\big(\|\mathbf{w}\|^2 - 1\big) + \mathbf{k}^\mathsf{T}(\bar{\mathbf{x}} - \mathbf{m}). \quad (5.2.8)$$

Lemma 5.2.4. *There exist $\mathbf{m}_N^2, \lambda_N^2$ and $\mathbf{k}_N^2$ such that $\boldsymbol{\theta}_N^2 \overset{\text{def}}{=} (\mathbf{a}_N^2, \mathbf{m}_N^2, \lambda_N^2, \mathbf{k}_N^2)$ is a solution of the system $\frac{1}{N}\sum_{t=1}^{N} \boldsymbol{\psi}^2(\mathbf{x}, \boldsymbol{\theta}) = 0$, where $\boldsymbol{\theta} = (\mathbf{w}, \mathbf{m}, \lambda, \mathbf{k})$ and*

$$\boldsymbol{\psi}^2(\mathbf{x}, \boldsymbol{\theta}) \overset{\text{def}}{=} \begin{bmatrix} g\Big(\mathbf{w}^\mathsf{T}(\mathbf{x} - \mathbf{m})\Big)(\mathbf{x} - \mathbf{m}) + 2\lambda\mathbf{w} \\ g\Big(\mathbf{w}^\mathsf{T}(\mathbf{x} - \mathbf{m})\Big)\mathbf{w} + \mathbf{k} \\ \|\mathbf{w}\|^2 - 1 \\ \bar{\mathbf{x}} - \mathbf{m} \end{bmatrix}.$$

Moreover, we have explicitly $\mathbf{m}_N^2 = \bar{\mathbf{x}}$ and

$$\lambda_N^2 = -\frac{1}{2N}\sum_{t=1}^{N} g\Big((\mathbf{a}_N^2)^\mathsf{T}(\mathbf{x}(t) - \bar{\mathbf{x}})\Big)(\mathbf{a}_N^2)^\mathsf{T}(\mathbf{x}(t) - \bar{\mathbf{x}}) \quad (5.2.9)$$

$$\mathbf{k}_N^2 = -\frac{1}{N}\sum_{t=1}^{N} g\Big((\mathbf{a}_N^2)^\mathsf{T}(\mathbf{x}(t) - \bar{\mathbf{x}})\Big)\mathbf{a}_N^2 \quad (5.2.10)$$

Proof. *Since $(\mathbf{a}_N^2, \bar{\mathbf{x}})$ is a solution of the optimization problem (5.2.7), it follows from Theorem 5.2.1 the existence of λ_N^2 and $\mathbf{k}_N^2$. Now it remains to*

prove (5.2.9) and (5.2.10). Note that $\frac{1}{N}\sum_{t=1}^{N}\psi^2(\mathbf{x},\boldsymbol{\theta})=0$ *is equivalent to*

$$\frac{1}{N}\sum_{t=1}^{N} g\Big(\mathbf{w}^{\mathsf{T}}(\mathbf{x}(t)-\mathbf{m})\Big)(\mathbf{x}-\mathbf{m})+2\lambda\mathbf{w}=0, \tag{5.2.11}$$

$$\frac{1}{N}\sum_{t=1}^{N} g\Big(\mathbf{w}^{\mathsf{T}}(\mathbf{x}(t)-\mathbf{m})\Big)\mathbf{w}+\mathbf{k}=0, \tag{5.2.12}$$

$$\|\mathbf{w}\|^2-1=0,$$

$$\bar{\mathbf{x}}-\mathbf{m}=0.$$

In (5.2.11), multiplying $\mathbf{w}^{\mathsf{T}}$ *on the left and substituting* $(\mathbf{w},\mathbf{m})$ *by* $(\mathbf{a}_N^2,\bar{\mathbf{x}})$*, we obtain (5.2.9) and (5.2.10) follows directly from (5.2.12).* □

Lagrange function for the case $k=3$.
We have

$$\mathbf{a}_N^3=\underset{\mathbf{w}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}}{\operatorname{argmin}} G(\mathbf{w},\mu_N^3)=\underset{\mathbf{w}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}}{\operatorname{argmin}}\frac{1}{N}\sum_{t=1}^{N}G\big(\mathbf{w}^{\mathsf{T}}\mathbf{Q}_N^{-1/2}\mathbf{x}(t)\big).$$

Here, $\Phi(\mathbf{w})=G(\mathbf{w},\mu_N^2)$ and $\mathbf{q}(\mathbf{w})=\|\mathbf{w}\|^2-1$. However, we prefer to write

$$\mathbf{a}_N^3=\mathbf{Q}_N^{1/2}\Big(\underset{\mathbf{w}^{\mathsf{T}}\mathbf{Q}_N^{1/2}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}}{\operatorname{argmin}}\frac{1}{N}\sum_{t=1}^{N}G\big(\mathbf{w}^{\mathsf{T}}\mathbf{x}(t)\big)\Big)$$

and study

$$\mathbf{b}_N^3\overset{\text{def}}{=}\underset{\mathbf{w}^{\mathsf{T}}\mathbf{Q}_N^{1/2}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}}{\operatorname{argmin}}\frac{1}{N}\sum_{t=1}^{N}G\big(\mathbf{w}^{\mathsf{T}}\mathbf{x}(t)\big) \tag{5.2.13}$$

for reasons that will be given in the next section. We note that $\hat{\mathbf{b}}_N^3=\mathbf{Q}_N^{-1/2}\mathbf{a}_N^3$ and the constraint appeared in (5.2.13) has the following equivalent form:

$$\{\mathbf{w}:\mathbf{w}^{\mathsf{T}}\mathbf{Q}_N^{1/2}\in\mathcal{S}\}=\{\mathbf{w}:\quad\mathbf{w}^{\mathsf{T}}\mathbf{Q}_N\mathbf{w}=1\}$$

It follows that the Lagrange function is given by

$$\mathscr{L}_3(\mathbf{w},\lambda)\overset{\text{def}}{=}\frac{1}{N}\sum_{t=1}^{N}G(\mathbf{w}^{\mathsf{T}}\mathbf{x}(t))+\lambda\big(\mathbf{w}^{\mathsf{T}}\mathbf{Q}_N\mathbf{w}-1\big).$$

Using the same proof as that of Lemma 5.2.2, we get the following result.

Lemma 5.2.5. *There exists* λ_N^3 *such that* $(\mathbf{b}_N^3, \lambda_N^3)$ *is a solution of the equation* $\frac{1}{N}\sum_{t=1}^{N} \boldsymbol{\psi}^3(\mathbf{x}, \boldsymbol{\theta}) = 0$, *where* $\boldsymbol{\theta} = (\mathbf{w}, \lambda)$ *and*

$$\boldsymbol{\psi}^3(\mathbf{x}, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \begin{bmatrix} g(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{x} + 2\lambda\mathbf{x}\mathbf{x}^\mathsf{T}\mathbf{w} \\ (\mathbf{w}^\mathsf{T}\mathbf{x})^2 - 1 \end{bmatrix}.$$

Moreover, we have explicitly

$$\lambda_N^3 = -\frac{1}{2N}\sum_{t=1}^{N} g\Big((\mathbf{b}_N^3)^\mathsf{T}\mathbf{x}(t)\Big)(\mathbf{b}_N^3)^\mathsf{T}\mathbf{x}(t).$$

Lagrange function for the case $k = 4$.
We have

$$\begin{aligned} \mathbf{a}_N^4 &= \operatorname*{argmin}_{\mathbf{w}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}} G(\mathbf{w}, \mu_N^k) \\ &= \operatorname*{argmin}_{\mathbf{w}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}} \frac{1}{N}\sum_{t=1}^{N} G\big(\mathbf{w}^\mathsf{T}\mathbf{C}_N^{-1/2}(\mathbf{x}(t) - \bar{\mathbf{x}})\big). \end{aligned} \tag{5.2.14}$$

For reasons that will be given in the next section, we prefer to write

$$\mathbf{a}_N^4 = \mathbf{C}_N^{1/2}\Big(\operatorname*{argmin}_{\mathbf{w}^\mathsf{T}\mathbf{C}_N^{1/2}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}} \frac{1}{N}\sum_{t=1}^{N} G\big(\mathbf{w}^\mathsf{T}(\mathbf{x}(t) - \bar{\mathbf{x}})\big)\Big) \tag{5.2.15}$$

and study

$$\mathbf{b}_N^4 \stackrel{\text{def}}{=} \operatorname*{argmin}_{\mathbf{w}^\mathsf{T}\mathbf{C}_N^{1/2}\in\mathscr{B}_r(\mathbf{a})\cap\mathcal{S}} \frac{1}{N}\sum_{t=1}^{N} G\big(\mathbf{w}^\mathsf{T}(\mathbf{x}(t) - \bar{\mathbf{x}})\big). \tag{5.2.16}$$

Then by (5.2.15) we have $\mathbf{b}_N^4 = \mathbf{C}_N^{-1/2}\mathbf{a}_N^k$. We note that the constraint appeared in (5.2.16) has the following equivalent form:

$$\begin{aligned} \{\mathbf{w} : \mathbf{w}^\mathsf{T}\mathbf{C}_N^{1/2} \in \mathcal{S}\} &= \{\mathbf{w} : \quad \mathbf{w}^\mathsf{T}\mathbf{C}_N\mathbf{w} = 1\} \\ &= \{\mathbf{w} : \quad \mathbf{w}^\mathsf{T}\mathbf{Q}_N\mathbf{w} - (\mathbf{w}^\mathsf{T}\bar{\mathbf{x}})^2 = 1\}. \end{aligned} \tag{5.2.17}$$

Introducing the auxiliary constraint $\bar{\mathbf{x}} = \mathbf{m}$, and in view of (5.2.17), we can write the optimization problem (5.2.16) in the following form:

$$\begin{aligned} \text{minimize} \quad & \Phi(\mathbf{w}, \mathbf{m}) = \frac{1}{N}\sum_{t=1}^{N} G\big(\mathbf{w}^\mathsf{T}(\mathbf{x}(t) - \mathbf{m})\big) \\ \text{subject to} \quad & \mathbf{q}(\mathbf{w}, \mathbf{m}) = \begin{bmatrix} \mathbf{w}^\mathsf{T}\mathbf{Q}_N\mathbf{w} - (\mathbf{w}^\mathsf{T}\mathbf{m})^2 - 1 \\ \bar{\mathbf{x}} - \mathbf{m} \end{bmatrix} = \mathbf{0}. \end{aligned}$$

It follows that the corresponding Lagrange function is given by

$$\mathscr{L}_4(\mathbf{w},\mathbf{m},\lambda,\mathbf{k}) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{t=1}^{N} G(\mathbf{w}^\mathsf{T}(\mathbf{x}(t)-\mathbf{m})) + \lambda\big(\mathbf{w}^\mathsf{T}\mathbf{Q}_N\mathbf{w} - (\mathbf{w}^\mathsf{T}\mathbf{m})^2 - 1\big)$$

$$+\mathbf{k}^\mathsf{T}(\bar{\mathbf{x}}-\mathbf{m}). \tag{5.2.18}$$

Using the same proof as that of Lemma 5.2.4, we get the following result.

Lemma 5.2.6. *There exist* $\mathbf{m}_N^4, \lambda_N^4$ *and* $\mathbf{k}_N^4$ *such that* $\boldsymbol{\theta}_N^4 \stackrel{\text{def}}{=} (\mathbf{b}_N^4, \mathbf{m}_N^4, \lambda_N^4, \mathbf{k}_N^4)$ *is a solution of the equation* $\frac{1}{N}\sum_{t=1}^{N}\boldsymbol{\psi}^4(\mathbf{x},\boldsymbol{\theta}) = 0$, *where* $\boldsymbol{\theta} = (\mathbf{w},\mathbf{m},\lambda,\mathbf{k})$ *and*

$$\boldsymbol{\psi}^4(\mathbf{x},\boldsymbol{\theta}) \stackrel{\text{def}}{=} \begin{bmatrix} g\big(\mathbf{w}^\mathsf{T}(\mathbf{x}-\mathbf{m})\big)(\mathbf{x}-\mathbf{m}) + 2\lambda\mathbf{x}\mathbf{x}^\mathsf{T}\mathbf{w} - 2\lambda(\mathbf{w}^\mathsf{T}\mathbf{m})\mathbf{m}, \\ (\mathbf{w}^\mathsf{T}\mathbf{x})^2 - (\mathbf{w}^\mathsf{T}\mathbf{m})^2 - 1, \\ \mathbf{x}-\mathbf{m}, \\ -g\big(\mathbf{w}^\mathsf{T}(\mathbf{x}-\mathbf{m})\big)\mathbf{w} - 2\lambda(\mathbf{w}^\mathsf{T}\mathbf{m})\mathbf{w} - \mathbf{k}. \end{bmatrix}.$$

Moreover, we have explicitly $\mathbf{m}_N^4 = \bar{\mathbf{x}}$ *and*

$$\lambda_N^4 = -\frac{1}{2N}\sum_{t=1}^{N} g\Big((\mathbf{b}_N^4)^\mathsf{T}(\mathbf{x}(t)-\mathbf{m}_N)\Big)(\mathbf{b}_N^4)^\mathsf{T}(\mathbf{x}(t)-\mathbf{m}_N)$$

$$\mathbf{k}_N^4 = -\frac{1}{N}\sum_{t=1}^{N} g\Big((\mathbf{b}_N^4)^\mathsf{T}(\mathbf{x}(t)-\mathbf{m}_N)\Big)\mathbf{b}_N^4 - 2\lambda_N^4((\mathbf{b}_N^4)^\mathsf{T}\mathbf{m}_N)\mathbf{b}_N^4,$$

5.3 M-estimators

Definition 5.3.1. *Let* $\boldsymbol{\psi} : \mathbb{R}^d \times \mathbb{R}^m \mapsto \mathbb{R}^n$ *be a measurable function. A solution* $\hat{\boldsymbol{\theta}}_N$ *of the equation*

$$\sum_{t=1}^{N}\boldsymbol{\psi}(\mathbf{x}(t),\boldsymbol{\theta}) = 0 \tag{5.3.1}$$

is called M-estimator.

We note that $\sum_{t=1}^{N}\boldsymbol{\psi}(\mathbf{x}(t),\boldsymbol{\theta})$ is an empirical approximation of $\mathbb{E}[\boldsymbol{\psi}(\mathbf{x}(t),\boldsymbol{\theta})]$ for all $\boldsymbol{\theta}$. Therefore, intuitively, under appropriate conditions the M-estimator $\hat{\boldsymbol{\theta}}_N$ obtained by solving (5.3.1) should converge to $\boldsymbol{\theta}_0$, which is a root of the equation $\mathbb{E}[\boldsymbol{\psi}(\mathbf{x}(t),\boldsymbol{\theta})] = 0$. This is indeed true. For a detailed discussion of this subject, we refer to (van der Vaart, 2000). Now that $\hat{\boldsymbol{\theta}}_N \to \boldsymbol{\theta}_0$, we are interested in the order at which the discrepancy $\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0$ converges to zero. Regarding this problem, we have the following result:

Theorem 5.3.2. *For each $\boldsymbol{\theta}$ in an open set of Euclidean space, let $\mathbf{x} \to \boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta})$ be a measurable vector valued function such that, for every $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in a neighborhood of $\boldsymbol{\theta}_0$ and a measurable function $K(\cdot)$ with $\mathbb{E}[K(\mathbf{x})^2] < \infty$,*

$$\|\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta}_1) - \boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta}_2)\| \leq K(\mathbf{x})\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

Assume that $\mathbb{E}[\|\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta}_0)\|^2] < \infty$ and that the map $\boldsymbol{\theta} \to \mathbb{E}[\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta})]$ is differentiable at a zero $\boldsymbol{\theta}_0$, with nonsingular derivative matrix $\mathbf{V}_{\boldsymbol{\theta}_0}$. If $\hat{\boldsymbol{\theta}}_N \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_0$, then

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \xrightarrow[N\to\infty]{\mathscr{D}} \mathbf{Z},$$

where $\mathbf{Z}$ is the solution of the system $\mathbf{V}_{\boldsymbol{\theta}_0}\mathbf{Z} = \mathbf{Y}$ with $\mathbf{Y} \sim \mathcal{N}(0, \mathbb{E}[\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta}_0)\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta}_0)^\mathsf{T}])$.

In order to apply the theorem, we need to find the function $\boldsymbol{\psi}$ associated to the optimization problem (5.2.3) for $k = 1, 2, 3, 4$, such that the corresponding zero satisfies $\hat{\boldsymbol{\theta}}_N \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_0$.

Proposition 5.3.3. *Let $\boldsymbol{\psi}_k(\mathbf{x}, \boldsymbol{\theta})$ and $\boldsymbol{\theta}_N^k$ be defined in Lemma 5.2.2, 5.2.4, 5.2.5, and 5.2.6 for $k = 1, 2, 3, 4$.*

(i). $\boldsymbol{\theta}_N^k \xrightarrow[N\to\infty]{a.s.} (\mathbf{a}, \lambda_\mathbf{a})$ *for* $k = 1, 3$,

(ii). $\boldsymbol{\theta}_N^k \xrightarrow[N\to\infty]{a.s.} (\mathbf{a}, \mathbf{m}_\mathbf{a}, \lambda_\mathbf{a}, \mathbf{k}_\mathbf{a})$ *for* $k = 2, 4$,

where $\mathbf{m}_\mathbf{a} = 0$, $\lambda_\mathbf{a} = -\frac{1}{2}\mathbb{E}[g(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a}^\mathsf{T}\mathbf{x}]$ and $\mathbf{k}_\mathbf{a} = -\mathbb{E}[g(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a}]$.

Proof. It suffices to use the same argument as in Lemma 4.2.4. □

Lemma 5.3.4. *There exists a neighborhood $\mathscr{B}_r(\boldsymbol{\theta}_\mathbf{a})$ of $\boldsymbol{\theta}_\mathbf{a}$, and a measurable function $K(\mathbf{x})$ with $\mathbb{E}[K(\mathbf{x})^2] < \infty$ such that for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathscr{B}_r(\boldsymbol{\theta}_\mathbf{a})$, we have*

$$\|\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta}_1) - \boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta}_2)\| \leq K(\mathbf{x})\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

Proof. See Appendix.

5.4 Numerical results

Example 5.4.1. The purpose of this example is to compare the theoretical value of Trace(Cov($\boldsymbol{z}_k$)) with its finite sample estimate. Our simulation is implemented as follows. We take a sample of size $N = 1000 \times 2^l$, with $l = 1, \ldots, 7$ corresponding to the horizonal axis in the figure. For each l, we run empirical FastICA to obtain the estimator $\mathbf{a}_N^k$, and then calculate $\varepsilon \overset{\text{def}}{=} N\|\mathbf{a}_N^k - \mathbf{a}\|^2$. We repeat this procedure independently 100 times, and

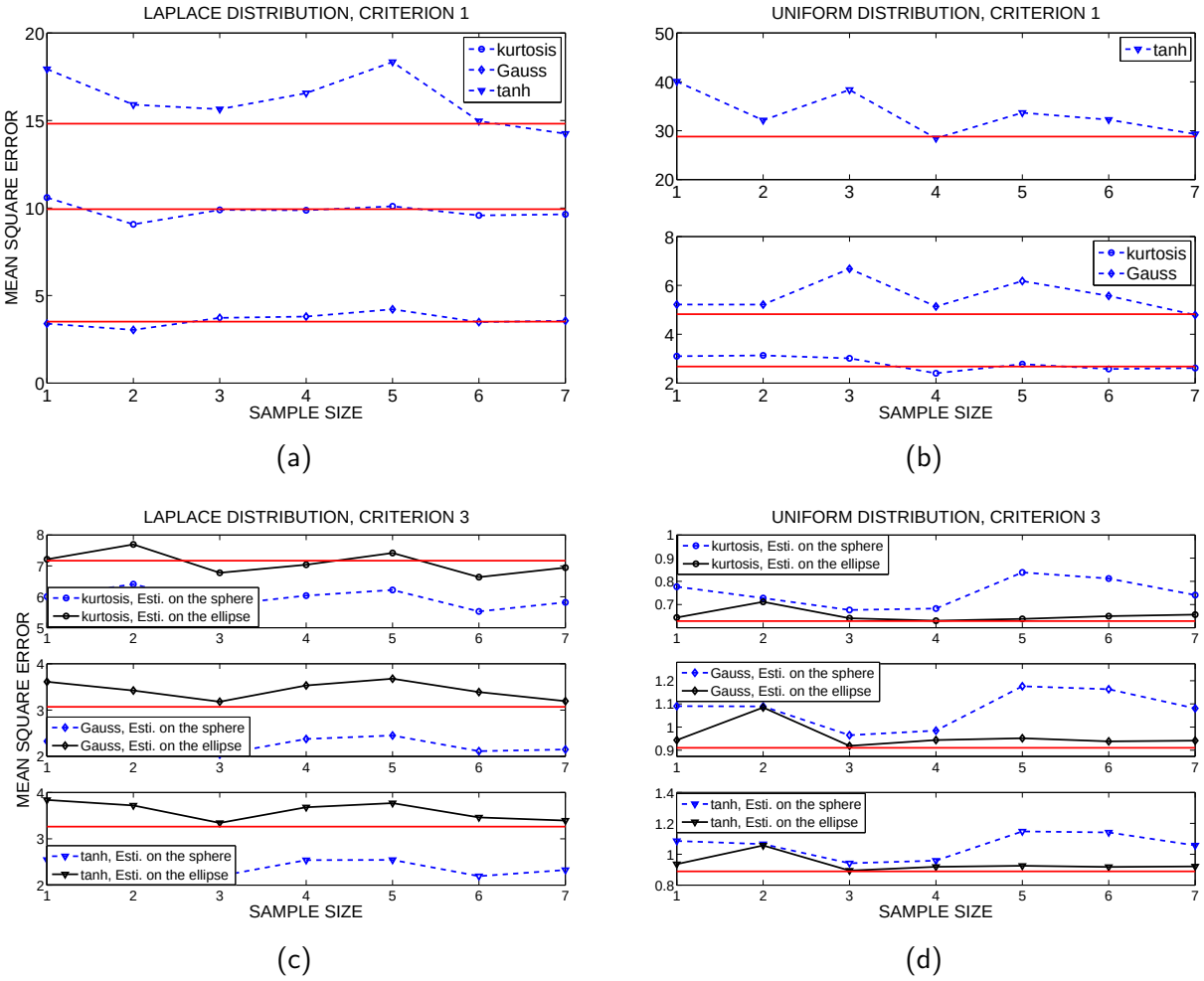

Figure 5.1: Mean square errors of $\mathbb{E}[N\|\mathbf{a}_N^1 - \mathbf{a}\|^2]$, $\mathbb{E}[N\|\mathbf{a}_N^3 - \mathbf{a}\|^2]$, $\mathbb{E}[N\|\mathbf{Q}_N^{-1/2}\mathbf{a}_N^3 - \mathbf{a}\|^2]$ versus the sample size $N = 1000 \times 2^n$.

take $\bar{\varepsilon} \overset{\text{def}}{=} \frac{1}{100}\sum_{t=1}^{100}\varepsilon(t)$ as the estimate of the Trace(Cov($\boldsymbol{z}_k$)). Letting l vary from 1 to 7, we then get a zigzag representing the finite sample estimate of Trace(Cov($\boldsymbol{z}_k$)). Our result, a zigzag and a straight line (representing the theoretical value of Trace(Cov($\boldsymbol{z}_k$))) is plotted in Fig. 5.1, where (a)-(d) each corresponds to a different case.

Figure (a) corresponds to the extraction of a Laplace source signal using μ_N^1 with three different nonlinearity functions, namely "kurtosis", "Gauss" and "tanh". From the figure, we observe that all three zigzags, each corresponding to a different nonlinearity function, coincide well with the respective straight lines. Figure (b) is the same as Figure (a) except for the uniform source signal. In Figure (c) and (d), we tackle the distribution μ_N^3. Although

this case concerns the estimator $\mathbf{Q}_N^{-1/2}\mathbf{a}_N^3$ rather than $\mathbf{a}_N^3$, as is shown in (5.1.3) and (5.1.7), in the simulation we considered both. More precisely, we plotted respectively the average of $N\|\mathbf{a}_N^k-\mathbf{a}\|^2$ and $N\|\mathbf{Q}_N^{-1/2}\mathbf{a}_N^3-\mathbf{a}\|^2$ in 100 independent trials. Thus, we have two zigzags and a straight line in (c) and (d), with the blue zigzag corresponding to $\mathbf{a}_N^3$ and the black corresponding to $\mathbf{Q}_N^{-1/2}\mathbf{a}_N^3$. (In the legend, they are marked as "Esti. on the sphere" and "Esti. on the ellipse" since $\mathbf{a}_N^3$ always lies on the unit sphere while $\mathbf{Q}_N^{-1/2}\mathbf{a}_N^3$ is not.) From the figure, we observe that the asymptotic covariance of these two estimators seems different. Besides, the black zigzag coincides as expected with its theoretical straight line, both for the extraction of a Laplace source signal (Figure (c)) and that of a uniform signal (Figure (d)). The conclusion is that Fig. 5.1 validates our theoretical results established in Theorem 5.1.1. Note that although it is not given here, theoretical result concerning μ_2 and μ_4 are also validated in the simulation. We choose to omit this part because the result is completely similar. We recall that our three nonlinearity functions are odd, and both Laplace and uniform distributions are symmetric, then see Remark 5.1.2.

Example 5.4.2. Table 5.1 concerns the theoretical value of $\mathrm{Trace}(\mathrm{Cov}(\boldsymbol{z}_k))$ for $k=1,2,3,4$. In this example, we consider three usual nonlinearity functions, namely "kurtosis", "Gauss" and "tanh", as well as a family probability distributions including the uniform distribution, the Laplace distribution, the generalized Gaussian distribution $\mathrm{GG}(\alpha)$ and the Gaussian mixture distribution $\mathrm{GM}(p,m)$. From the table, we observe that the empirical FastICA algorithm with respect to μ_N^3 and μ_N^4 performs better in terms of asymptotic error (i.e. the corresponding trace of asymptotic covariance matrix is smaller). Besides, if the extracted source signal has a non-symmetrical distribution, then the term η_i in (5.1.8) does not vanish, resulting in $\mathrm{Trace}(\mathrm{Cov}(\boldsymbol{z}_4)) > \mathrm{Trace}(\mathrm{Cov}(\boldsymbol{z}_3))$. In this case, μ_N^3 is even more preferable than μ_N^4.

	uniform	Laplace	GM($\frac{1}{6}$, 2)	GM($\frac{1}{10}$, 2.5)	GG(0.5)	GG(3)
$k=1$	(kurt, 2.68)	(gaus, 3.51)	(gaus, 5.13)	(gaus, 1.90)	(kurt, 4.13)	(kurt, 25.3)
$k=2$	(kurt, 2.68)	(gaus, 3.51)	(gaus, 8.61)	(gaus, 2.71)	(kurt, 4.13)	(kurt, 25.3)
$k=3$	(kurt, 0.43)	(gaus, 1.82)	(gaus, 2.56)	(gaus, 1.15)	(gaus, 0.32)	(tanh, 7.73)
$k=4$	(kurt, 0.43)	(gaus, 1.82)	(gaus, 6.03)	(gaus, 1.97)	(gaus, 0.32)	(tanh, 7.73)

Table 5.1: Theoretical value of $\mathrm{Trace}(\mathrm{Cov}(\boldsymbol{z}_k))$ for different nonlinearity functions and different distribution of the source signal.

5.5 Proofs

5.5.1 Proof of Lemma 5.3.4

We recall that

$$\boldsymbol{\psi}(\mathbf{x},\boldsymbol{\theta}) = \begin{bmatrix} g\big(\mathbf{w}^\mathsf{T}(\mathbf{x}-\mathbf{m})\big)(\mathbf{x}-\mathbf{m}) + 2\lambda\mathbf{x}\mathbf{x}^\mathsf{T}\mathbf{w} - 2\lambda(\mathbf{w}^\mathsf{T}\mathbf{m})\mathbf{m} \\ -g\big(\mathbf{w}^\mathsf{T}(\mathbf{x}-\mathbf{m})\big)\mathbf{w} - 2\lambda(\mathbf{w}^\mathsf{T}\mathbf{m})\mathbf{w} - \mathbf{k} \\ (\mathbf{w}^\mathsf{T}\mathbf{x})^2 - (\mathbf{w}^\mathsf{T}\mathbf{m})^2 - 1 \\ \mathbf{x}-\mathbf{m} \end{bmatrix}.$$

By Assumption 3, the function g has continuous derivative and satisfies $g(t) \le c|t|^p$. It follows that

$$\boldsymbol{\psi}(\mathbf{x},\boldsymbol{\theta}_1) - \boldsymbol{\psi}(\mathbf{x},\boldsymbol{\theta}_2) = \partial_{\boldsymbol{\theta}}\boldsymbol{\psi}(\mathbf{x},\boldsymbol{\xi})(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2), \tag{5.5.1}$$

where each component of the matrix $\partial_{\boldsymbol{\theta}}\boldsymbol{\psi}(\mathbf{x},\boldsymbol{\xi})$ can be bounded by a polynomial of $\|\mathbf{x}\|$ and $\|\boldsymbol{\xi}\|$. Since $\boldsymbol{\xi}$ is an intermediate point between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, which belongs to the compact $\mathscr{B}_r(\boldsymbol{\theta}_0)$, we can find a polynomial function $K(\cdot)$ such that

$$\|\partial_{\boldsymbol{\theta}}\boldsymbol{\psi}(\mathbf{x},\boldsymbol{\xi})\| \le K(\mathbf{x})$$

and $\mathbb{E}[K(\mathbf{x})^2] < +\infty$. Hence, we deduce from (5.5.1) that

$$\|\boldsymbol{\psi}(\mathbf{x},\boldsymbol{\theta}_1) - \boldsymbol{\psi}(\mathbf{x},\boldsymbol{\theta}_2)\| \le K(\mathbf{x})\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

5.5.2 Proof of Theorem 5.1.1

Here we give only the proof for the case $k = 4$ since it is the most complicated case, and the same method applies to the cases $k = 1, 2, 3$. In the sequel, to simplify the notation, we will omit the superscript of $\boldsymbol{\psi}^4$ and write $\boldsymbol{\psi}(\mathbf{x},\boldsymbol{\theta})$ instead. Besides, we denote $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3,\boldsymbol{\theta}_4) = (\mathbf{w},\mathbf{m},\lambda,\mathbf{k})$,

$$\boldsymbol{\psi}(\mathbf{x},\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\psi}_1(\mathbf{x},\boldsymbol{\theta}) \\ \boldsymbol{\psi}_2(\mathbf{x},\boldsymbol{\theta}) \\ \boldsymbol{\psi}_3(\mathbf{x},\boldsymbol{\theta}) \\ \boldsymbol{\psi}_4(\mathbf{x},\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} g\big(\mathbf{w}^\mathsf{T}(\mathbf{x}-\mathbf{m})\big)(\mathbf{x}-\mathbf{m}) + 2\lambda\mathbf{x}\mathbf{x}^\mathsf{T}\mathbf{w} - 2\lambda(\mathbf{w}^\mathsf{T}\mathbf{m})\mathbf{m} \\ -g\big(\mathbf{w}^\mathsf{T}(\mathbf{x}-\mathbf{m})\big)\mathbf{w} - 2\lambda(\mathbf{w}^\mathsf{T}\mathbf{m})\mathbf{w} - \mathbf{k} \\ (\mathbf{w}^\mathsf{T}\mathbf{x})^2 - (\mathbf{w}^\mathsf{T}\mathbf{m})^2 - 1 \\ \mathbf{x}-\mathbf{m} \end{bmatrix}.$$

and

$$\partial_{\boldsymbol{\theta}}\boldsymbol{\psi}(\mathbf{x},\boldsymbol{\theta}) \stackrel{\text{def}}{=} \begin{bmatrix} \boldsymbol{\psi}_{11} & \boldsymbol{\psi}_{12} & \boldsymbol{\psi}_{13} & \boldsymbol{\psi}_{14} \\ \boldsymbol{\psi}_{21} & \boldsymbol{\psi}_{22} & \boldsymbol{\psi}_{23} & \boldsymbol{\psi}_{24} \\ \boldsymbol{\psi}_{31} & \boldsymbol{\psi}_{32} & \boldsymbol{\psi}_{33} & \boldsymbol{\psi}_{34} \\ \boldsymbol{\psi}_{41} & \boldsymbol{\psi}_{42} & \boldsymbol{\psi}_{43} & \boldsymbol{\psi}_{44} \end{bmatrix},$$

where $\boldsymbol{\psi}_{ij} \stackrel{\text{def}}{=} \partial_{\boldsymbol{\theta}_j}\boldsymbol{\psi}_i$.

From Proposition 5.3.3 and Lemma 5.3.4, we see that the hypotheses of Theorem 5.3.2 are satisfied for $\hat{\boldsymbol{\theta}}_N, \boldsymbol{\theta}_{\mathbf{a}}$ and $\boldsymbol{\psi}$. Applying Theorem 5.3.2, we get the following asymptotic result:

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_{\mathbf{a}}) \xrightarrow[N\to\infty]{\mathscr{D}} \mathbf{Z}, \tag{5.5.2}$$

where $\mathbf{Z}$ is such that

$$\mathbb{E}[\partial_{\boldsymbol{\theta}}\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta}_{\mathbf{a}})]\mathbf{Z} = \mathbf{Y} \tag{5.5.3}$$

with $\mathbf{Y} \sim \mathcal{N}(0, \mathbb{E}[\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta}_{\mathbf{a}})\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta}_{\mathbf{a}})^{\mathsf{T}}])$. Let us denote

$$\mathbf{P} = \mathbb{E}[\partial_{\boldsymbol{\theta}}\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta}_{\mathbf{a}})] = \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \mathbf{P}_{13} & \mathbf{P}_{14} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \mathbf{P}_{23} & \mathbf{P}_{24} \\ \mathbf{P}_{31} & \mathbf{P}_{32} & \mathbf{P}_{33} & \mathbf{P}_{34} \\ \mathbf{P}_{41} & \mathbf{P}_{42} & \mathbf{P}_{43} & \mathbf{P}_{44} \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} \boldsymbol{z}_{\mathbf{a}} \\ \boldsymbol{z}_{\mathbf{m}} \\ \boldsymbol{z}_{\lambda} \\ \boldsymbol{z}_{\mathbf{k}} \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} \boldsymbol{y}_{\mathbf{a}} \\ \boldsymbol{y}_{\mathbf{m}} \\ \boldsymbol{y}_{\lambda} \\ \boldsymbol{y}_{\mathbf{k}} \end{bmatrix}.$$

Then system (5.5.3) can be written as

$$\begin{aligned} \boldsymbol{y}_{\mathbf{a}} &= \mathbf{P}_{11}\boldsymbol{z}_{\mathbf{a}} + \mathbf{P}_{12}\boldsymbol{z}_{\mathbf{m}} + \mathbf{P}_{13}\boldsymbol{z}_{\lambda} + \mathbf{P}_{14}\boldsymbol{z}_{\mathbf{k}}, \\ \boldsymbol{y}_{\mathbf{m}} &= \mathbf{P}_{21}\boldsymbol{z}_{\mathbf{a}} + \mathbf{P}_{22}\boldsymbol{z}_{\mathbf{m}} + \mathbf{P}_{23}\boldsymbol{z}_{\lambda} + \mathbf{P}_{24}\boldsymbol{z}_{\mathbf{k}}, \\ \boldsymbol{y}_{\lambda} &= \mathbf{P}_{31}\boldsymbol{z}_{\mathbf{a}} + \mathbf{P}_{32}\boldsymbol{z}_{\mathbf{m}} + \mathbf{P}_{33}\boldsymbol{z}_{\lambda} + \mathbf{P}_{34}\boldsymbol{z}_{\mathbf{k}}, \\ \boldsymbol{y}_{\mathbf{k}} &= \mathbf{P}_{41}\boldsymbol{z}_{\mathbf{a}} + \mathbf{P}_{42}\boldsymbol{z}_{\mathbf{m}} + \mathbf{P}_{43}\boldsymbol{z}_{\lambda} + \mathbf{P}_{44}\boldsymbol{z}_{\mathbf{k}}. \end{aligned}$$

Note that we are only interested in $\boldsymbol{z}_{\mathbf{a}}$. In what follows, we will compute the explicit form of $\boldsymbol{z}_{\mathbf{a}}$ and deduce its covariance matrix.

Step 1. In the first step, we will compute the matrix $\mathbb{E}[\partial_{\boldsymbol{\theta}}\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta}_{\mathbf{a}})]$ and $\mathbb{E}[\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta}_{\mathbf{a}})\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta}_{\mathbf{a}})^{\mathsf{T}}]$ explicitly. Let us denote

$$\mathbf{R} = \mathbb{E}[\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta}_{\mathbf{a}})\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta}_{\mathbf{a}})^{\mathsf{T}}] = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} & \mathbf{R}_{13} & \mathbf{R}_{14} \\ \mathbf{R}_{21} & \mathbf{R}_{22} & \mathbf{R}_{23} & \mathbf{R}_{24} \\ \mathbf{R}_{31} & \mathbf{R}_{32} & \mathbf{R}_{33} & \mathbf{R}_{34} \\ \mathbf{R}_{41} & \mathbf{R}_{42} & \mathbf{R}_{43} & \mathbf{R}_{44} \end{bmatrix}.$$

Since $\mathbf{P}, \mathbf{R}$ and $\partial_{\boldsymbol{\theta}}\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta})$ are all symmetric matrices, it suffices to calculate their upper triangular elements. We have

$$\begin{aligned} \boldsymbol{\psi}_{11} &= g'\big(\mathbf{w}^{\mathsf{T}}(\mathbf{x}-\mathbf{m})\big)(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^{\mathsf{T}} + 2\lambda(\mathbf{x}\mathbf{x}^{\mathsf{T}} - \mathbf{m}\mathbf{m}^{\mathsf{T}}), \\ \boldsymbol{\psi}_{12} &= -g'(\mathbf{w}^{\mathsf{T}}(\mathbf{x}-\mathbf{m}))(\mathbf{x}-\mathbf{m})\mathbf{w}^{\mathsf{T}} - g(\mathbf{w}^{\mathsf{T}}(\mathbf{x}-\mathbf{m}))\mathbf{I} - 2\big(\lambda\,\mathbf{m}\mathbf{w}^{\mathsf{T}} + (\mathbf{w}^{\mathsf{T}}\mathbf{m})\mathbf{I}\big), \\ \boldsymbol{\psi}_{13} &= 2(\mathbf{w}^{\mathsf{T}}\mathbf{x})\mathbf{x} - 2(\mathbf{w}^{\mathsf{T}}\mathbf{m})\mathbf{m}, \\ \boldsymbol{\psi}_{14} &= 0, \\ \boldsymbol{\psi}_{22} &= \big(g'(\mathbf{w}^{\mathsf{T}}(\mathbf{x}-\mathbf{m})) - 2\lambda\big)\mathbf{w}\mathbf{w}^{\mathsf{T}}, \\ \boldsymbol{\psi}_{23} &= -2(\mathbf{w}^{\mathsf{T}}\mathbf{m})\mathbf{w}^{\mathsf{T}} \\ \boldsymbol{\psi}_{24} &= -\mathbf{I}, \\ \boldsymbol{\psi}_{33} &= 0, \\ \boldsymbol{\psi}_{34} &= 0, \\ \boldsymbol{\psi}_{44} &= 0. \end{aligned}$$

Hence

$$\mathbf{P}_{11} = \mathbb{E}\big[\big(g'(\mathbf{a}^\mathsf{T}\mathbf{x}) + 2\lambda_\mathbf{a}\big)\mathbf{x}\mathbf{x}^\mathsf{T}\big], \tag{5.5.4}$$
$$\mathbf{P}_{12} = -\mathbb{E}[g'(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{x}\mathbf{a}^\mathsf{T}] - \mathbb{E}[g(\mathbf{a}^\mathsf{T}\mathbf{x})]\mathbf{I}, \tag{5.5.5}$$
$$\mathbf{P}_{13} = 2\mathbf{a}, \tag{5.5.6}$$
$$\mathbf{P}_{14} = 0, \tag{5.5.7}$$
$$\mathbf{P}_{22} = \mathbb{E}[g'(\mathbf{a}^\mathsf{T}\mathbf{x}) - 2\lambda_\mathbf{a}]\mathbf{a}\mathbf{a}^\mathsf{T}, \tag{5.5.8}$$
$$\mathbf{P}_{23} = 0, \tag{5.5.9}$$
$$\mathbf{P}_{24} = -\mathbf{I}, \tag{5.5.10}$$
$$\mathbf{P}_{33} = 0, \tag{5.5.11}$$
$$\mathbf{P}_{34} = 0, \tag{5.5.12}$$
$$\mathbf{P}_{44} = 0, \tag{5.5.13}$$
$$\tag{5.5.14}$$

To calculate $\mathbf{R}$, we first notice that

$$\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta}_\mathbf{a}) = \begin{bmatrix} -g(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a} - \mathbf{k}_\mathbf{a} \\ \big(g(\mathbf{a}^\mathsf{T}\mathbf{x}) + 2\lambda_\mathbf{a}(\mathbf{a}^\mathsf{T}\mathbf{x})\big)\mathbf{x} \\ (\mathbf{a}^\mathsf{T}\mathbf{x})^2 - 1 \\ \mathbf{x} \end{bmatrix}.$$

Then we have

$$\mathbf{R}_{11} = \mathbb{E}\big[\big(g(\mathbf{a}^\mathsf{T}\mathbf{x}) + 2\lambda_\mathbf{a}(\mathbf{a}^\mathsf{T}\mathbf{x})\big)^2\mathbf{x}\mathbf{x}^\mathsf{T}\big], \tag{5.5.15}$$
$$\mathbf{R}_{12} = -\mathbb{E}\big[\big(g(\mathbf{a}^\mathsf{T}\mathbf{x}) + 2\lambda_\mathbf{a}(\mathbf{a}^\mathsf{T}\mathbf{x})\big)(g(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a}^\mathsf{T} + \mathbf{k}_\mathbf{a}^\mathsf{T})\big], \tag{5.5.16}$$
$$\mathbf{R}_{13} = \mathbb{E}\big[\big(g(\mathbf{a}^\mathsf{T}\mathbf{x}) + 2\lambda_\mathbf{a}(\mathbf{a}^\mathsf{T}\mathbf{x})\big)((\mathbf{a}^\mathsf{T}\mathbf{x})^2 - 1)\mathbf{x}\big], \tag{5.5.17}$$
$$\mathbf{R}_{14} = \mathbb{E}\big[\big(g(\mathbf{a}^\mathsf{T}\mathbf{x}) + 2\lambda_\mathbf{a}(\mathbf{a}^\mathsf{T}\mathbf{x})\big)\mathbf{x}\mathbf{x}^\mathsf{T}\big], \tag{5.5.18}$$
$$\begin{aligned} \mathbf{R}_{22} &= \mathbb{E}[(g(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a} + \mathbf{k}_\mathbf{a})(g(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a}^\mathsf{T} + \mathbf{k}_\mathbf{a}^\mathsf{T})], \\ \mathbf{R}_{23} &= -\mathbb{E}\big[((\mathbf{a}^\mathsf{T}\mathbf{x})^2 - 1)(g(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a}^\mathsf{T} + \mathbf{k}_\mathbf{a}^\mathsf{T})\big], \\ \mathbf{R}_{24} &= -\mathbb{E}[(g(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{x}\mathbf{a}^\mathsf{T} + \mathbf{x}\mathbf{k}_\mathbf{a}^\mathsf{T})], \end{aligned} \tag{5.5.19}$$
$$\mathbf{R}_{33} = \mathbb{E}\big[((\mathbf{a}^\mathsf{T}\mathbf{x})^2 - 1)^2\big], \tag{5.5.20}$$
$$\mathbf{R}_{34} = \mathbb{E}\big[((\mathbf{a}^\mathsf{T}\mathbf{x})^2 - 1)\mathbf{x}^\mathsf{T}\big], \tag{5.5.21}$$
$$\mathbf{R}_{44} = \mathbf{I}. \tag{5.5.22}$$

Step 2. Let us denote

$$\mathbf{Z} = \begin{bmatrix} \boldsymbol{z}_\mathbf{a} \\ \boldsymbol{z}_\mathbf{m} \\ \boldsymbol{z}_\lambda \\ \boldsymbol{z}_\mathbf{k} \end{bmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{bmatrix} \boldsymbol{y}_\mathbf{a} \\ \boldsymbol{y}_\mathbf{m} \\ \boldsymbol{y}_\lambda \\ \boldsymbol{y}_\mathbf{k} \end{bmatrix}.$$

The first line of the system $\mathbf{PZ} = \mathbf{Y}$ gives

$$\begin{aligned}\boldsymbol{y}_{\mathbf{a}} &= \mathbf{P}_{11}\boldsymbol{z}_{\mathbf{a}} + \mathbf{P}_{12}\boldsymbol{z}_{\mathbf{m}} + \mathbf{P}_{13}\boldsymbol{z}_{\lambda} + \mathbf{P}_{14}\boldsymbol{z}_{\alpha} \\ &= \mathbb{E}\big[\big(g'(\mathbf{a}^{\mathsf{T}}\mathbf{x}) + 2\lambda\big)\mathbf{x}\mathbf{x}^{\mathsf{T}}\big]\boldsymbol{z}_{\mathbf{a}} - \mathbb{E}[g'(\mathbf{a}^{\mathsf{T}}\mathbf{x})\mathbf{x}\mathbf{a}^{\mathsf{T}} + g(\mathbf{a}^{\mathsf{T}}\mathbf{x})\mathbf{I}]\boldsymbol{z}_{\mathbf{m}} + 2\mathbf{a}\boldsymbol{z}_{\lambda}.\end{aligned}$$

Multiplying both sides of the equation above by $(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})$ gives

$$\begin{aligned}(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})\boldsymbol{y}_{\mathbf{a}} &= \mathbb{E}\big[\big(g'(\mathbf{a}^{\mathsf{T}}\mathbf{x}) + 2\lambda_{\mathbf{a}}\big)(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})\mathbf{x}\mathbf{x}^{\mathsf{T}}\big]\boldsymbol{z}_{\mathbf{a}} + 2(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})\mathbf{a}\boldsymbol{z}_{\lambda} \\ &\quad -\mathbb{E}[g'(\mathbf{a}^{\mathsf{T}}\mathbf{x})(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})\mathbf{x}\mathbf{a}^{\mathsf{T}} + g(\mathbf{a}^{\mathsf{T}}\mathbf{x})(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})]\boldsymbol{z}_{\mathbf{m}} \\ &= \mathbb{E}\big[\big(g'(\mathbf{a}^{\mathsf{T}}\mathbf{x}) - g(\mathbf{a}^{\mathsf{T}}\mathbf{x})\mathbf{a}^{\mathsf{T}}\mathbf{x}\big)\big](\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})\boldsymbol{z}_{\mathbf{a}} \\ &\quad -\mathbb{E}[g(\mathbf{a}^{\mathsf{T}}\mathbf{x})(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})]\boldsymbol{z}_{\mathbf{m}},\end{aligned}$$

or equivalently

$$\begin{aligned}&\frac{(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})\boldsymbol{y}_{\mathbf{a}}}{\mathbb{E}\big[\big(g'(\mathbf{a}^{\mathsf{T}}\mathbf{x}) - g(\mathbf{a}^{\mathsf{T}}\mathbf{x})\mathbf{a}^{\mathsf{T}}\mathbf{x}\big)\big]} \\ &= (\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})\boldsymbol{z}_{\mathbf{a}} - \frac{\mathbb{E}[g(\mathbf{a}^{\mathsf{T}}\mathbf{x})(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})]}{\mathbb{E}\big[\big(g'(\mathbf{a}^{\mathsf{T}}\mathbf{x}) - g(\mathbf{a}^{\mathsf{T}}\mathbf{x})\mathbf{a}^{\mathsf{T}}\mathbf{x}\big)\big]}\boldsymbol{z}_{\mathbf{m}}. \qquad (5.5.23)\end{aligned}$$

Besides, by (5.5.5)-(5.5.12) we have

$$\begin{aligned}\boldsymbol{y}_{\mathbf{m}} &= \mathbf{P}_{21}\boldsymbol{z}_{\mathbf{a}} + \mathbf{P}_{22}\boldsymbol{z}_{\mathbf{m}} + \mathbf{P}_{23}\boldsymbol{z}_{\lambda} + \mathbf{P}_{24}\boldsymbol{z}_{\mathbf{k}} = 2\mathbf{a}^{\mathsf{T}}\boldsymbol{z}_{\mathbf{a}}, \qquad (5.5.24) \\ \boldsymbol{y}_{\lambda} &= \mathbf{P}_{31}\boldsymbol{z}_{\mathbf{a}} + \mathbf{P}_{32}\boldsymbol{z}_{\mathbf{m}} + \mathbf{P}_{33}\boldsymbol{z}_{\lambda} + \mathbf{P}_{34}\boldsymbol{z}_{\mathbf{k}} = -\boldsymbol{z}_{\mathbf{m}},\end{aligned}$$

among which (5.5.24) implies $\mathbf{a}\boldsymbol{y}_{\lambda}/2 = \mathbf{a}\mathbf{a}^{\mathsf{T}}\boldsymbol{z}_{\mathbf{a}}$.

Using these results to (5.5.23) gives

$$\boldsymbol{z}_{\mathbf{a}} = \frac{(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})\boldsymbol{y}_{\mathbf{a}}}{\mathbb{E}\big[\big(g'(\mathbf{a}^{\mathsf{T}}\mathbf{x}) - g(\mathbf{a}^{\mathsf{T}}\mathbf{x})\mathbf{a}^{\mathsf{T}}\mathbf{x}\big)\big]} + \frac{\mathbf{a}\boldsymbol{y}_{\lambda}}{2} + \frac{\mathbb{E}[g(\mathbf{a}^{\mathsf{T}}\mathbf{x})](\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})\boldsymbol{y}_{\mathbf{k}}}{\mathbb{E}\big[\big(g'(\mathbf{a}^{\mathsf{T}}\mathbf{x}) - g(\mathbf{a}^{\mathsf{T}}\mathbf{x})\mathbf{a}^{\mathsf{T}}\mathbf{x}\big)\big]}.$$

Step 3. In the last step, we will use the fact that $\mathrm{Cov}(\mathbf{Y}) = \mathbf{R}$ to compute $\mathrm{Cov}(\boldsymbol{z}_{\mathbf{a}})$. Note that $H(\mathbf{a}, \mu) = \mathbb{E}\big[g'(\mathbf{a}^{\mathsf{T}}\mathbf{x}) - g(\mathbf{a}^{\mathsf{T}}\mathbf{x})\mathbf{a}^{\mathsf{T}}\mathbf{x}\big]$. Now for simplicity we write

$$\boldsymbol{z}_{\mathbf{a}} = c_1(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})\boldsymbol{y}_{\mathbf{a}} + c_3\mathbf{a}\boldsymbol{y}_{\lambda} + c_4(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})\boldsymbol{y}_{\mathbf{k}},$$

where

$$c_1 = \frac{1}{H(\mathbf{a}, \mu)}, \quad c_3 = \frac{1}{2}, \quad c_4 = \frac{\mathbb{E}[g(\mathbf{a}^{\mathsf{T}}\mathbf{x})]}{H(\mathbf{a}, \mu)}.$$

We notice that (5.5.2) yields $\mathrm{Cov}(\mathbf{Y}) = \mathbf{R}$. Thus

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{z}_{\mathbf{a}}) = {} & c_1^2(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})\mathbf{R}_{11}(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}}) + c_3^2\mathbf{a}\mathbf{R}_{33}\mathbf{a}^{\mathsf{T}} + c_4^2(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})\mathbf{R}_{44}(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}}) \\ & + c_1c_3(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})\mathbf{R}_{13}\mathbf{a}^{\mathsf{T}} + c_1c_3\mathbf{a}\mathbf{R}_{31}(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}}) + c_1c_4(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})\mathbf{R}_{14}(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}}) \\ & + c_1c_4(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})\mathbf{R}_{41}(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}}) + c_3c_4\mathbf{a}\mathbf{R}_{34}(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}}) + c_3c_4(\mathbf{I} - \mathbf{a}\mathbf{a}^{\mathsf{T}})\mathbf{R}_{34}\mathbf{a}^{\mathsf{T}}.\end{aligned}$$

Using (5.5.15)-(5.5.20), we deduce that

$$\begin{aligned}
(\mathbf{I}-\mathbf{a}\mathbf{a}^{\mathsf{T}})\mathbf{R}_{11}(\mathbf{I}-\mathbf{a}\mathbf{a}^{\mathsf{T}}) &= \mathbb{E}\big[\big(g(\mathbf{a}^{\mathsf{T}}\mathbf{x})+2\lambda_{\mathbf{a}}(\mathbf{a}^{\mathsf{T}}\mathbf{x})\big)^2\big](\mathbf{I}-\mathbf{a}\mathbf{a}^{\mathsf{T}}) \\
&= \Big(\mathbb{E}\big[\big(g(\mathbf{a}^{\mathsf{T}}\mathbf{x})\big)^2\big]-\big(\mathbb{E}\big[g(\mathbf{a}^{\mathsf{T}}\mathbf{x})\mathbf{a}^{\mathsf{T}}\mathbf{x}\big]\big)^2\Big)(\mathbf{I}-\mathbf{a}\mathbf{a}^{\mathsf{T}}), \\
\mathbf{a}\mathbf{R}_{33}\mathbf{a}^{\mathsf{T}} &= \mathrm{Cov}\big((\mathbf{a}^{\mathsf{T}}\mathbf{x})^2\big)(\mathbf{a}\mathbf{a}^{\mathsf{T}}), \\
(\mathbf{I}-\mathbf{a}\mathbf{a}^{\mathsf{T}})\mathbf{R}_{44}(\mathbf{I}-\mathbf{a}\mathbf{a}^{\mathsf{T}}) &= \mathbf{I}-\mathbf{a}\mathbf{a}^{\mathsf{T}}, \\
(\mathbf{I}-\mathbf{a}\mathbf{a}^{\mathsf{T}})\mathbf{R}_{13}\mathbf{a}^{\mathsf{T}} &= 0, \\
(\mathbf{I}-\mathbf{a}\mathbf{a}^{\mathsf{T}})\mathbf{R}_{14}(\mathbf{I}-\mathbf{a}\mathbf{a}^{\mathsf{T}}) &= \mathbb{E}\big[g(\mathbf{a}^{\mathsf{T}}\mathbf{x})\big](\mathbf{I}-\mathbf{a}\mathbf{a}^{\mathsf{T}}), \\
\mathbf{a}\mathbf{R}_{34}(\mathbf{I}-\mathbf{a}\mathbf{a}^{\mathsf{T}}) &= 0.
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{z}_4) = &\frac{\mathbb{E}\Big[\big(g(\mathbf{a}^{\mathsf{T}}\mathbf{x})\big)^2\Big]-\Big(\mathbb{E}\big[g(\mathbf{a}^{\mathsf{T}}\mathbf{x})\mathbf{a}^{\mathsf{T}}\mathbf{x}\big]\Big)^2+3\Big(\mathbb{E}[g(\mathbf{a}^{\mathsf{T}}\mathbf{x})]\Big)^2}{\Big(\mathbb{E}\big[\big(g'(\mathbf{a}^{\mathsf{T}}\mathbf{x})-g(\mathbf{a}^{\mathsf{T}}\mathbf{x})\mathbf{a}^{\mathsf{T}}\mathbf{x}\big)\big]\Big)^2}(\mathbf{I}-\mathbf{a}\mathbf{a}^{\mathsf{T}}) \\
&+\frac{\mathrm{Cov}\big((\mathbf{a}^{\mathsf{T}}\mathbf{x})^2\big)}{4}(\mathbf{a}\mathbf{a}^{\mathsf{T}}). \qquad (5.5.25)
\end{aligned}$$

□

Chapter 6

Asymptotic Analysis of the Gradient of the FastICA Function

Contents

In Chapter 3 (see e.g. Example 4.3.3), we have shown that although the theoretical FastICA converges with a quadratic convergence speed, the convergence speed of empirical FastICA is only linear. Clearly, the convergence speed, characterized by $\lim_{n\to\infty} \|\mathbf{w}_{n+1} - \mathbf{a}_N^k\|/\|\mathbf{w}_n - \mathbf{a}_N^k\|$, depends on the gradient $\nabla \boldsymbol{f}(\mathbf{a}_N^k, \mu_N^k)$ which is close to $\nabla \boldsymbol{f}(\mathbf{a}, \mu_N^k)$ for large N. Thus to study the convergence speed, it is better to focus on $\nabla \boldsymbol{f}(\mathbf{a}, \mu_N^k)$ since it is more analytically trackable than $\nabla \boldsymbol{f}(\mathbf{a}_N^k, \mu_N^k)$. A new criterion of optimality of the nonlinearity function, suggested by Hyvärinen to us, is to optimize the asymptotic covariance matrix of $N^{1/2}\nabla \boldsymbol{f}(\mathbf{a}, \mu_N^k)$, such that one can attain the fastest possible convergence speed of FastICA.

In this chapter, we will derive the explicit form of the latter asymptotic covariance matrix for $k = 1$ and give some numerical example for the two-dimensional case.

6.1 Statement of the main result

In the sequel, for simplicity we suppose that the mixing matrix $\mathbf{A}$ of model 3.1.1 is the identity matrix, and we are interested in the extraction of the source signal s_1, which in turn corresponds to the estimation of the first column of $\mathbf{A}$, namely, $\mathbf{e}_1$. Before announcing our main result, let us introduce some notations. We denote by $\mathrm{vec}(\cdot)$ the operation that reshapes columns of a matrix in one long column vector. For a random matrix $\mathbf{M}$, we write $\mathrm{Cov}(\mathbf{M}) \stackrel{\text{def}}{=} \mathrm{Cov}(\mathrm{vec}(\mathbf{M}))$. Besides, we define

$$\boldsymbol{h}(\mathbf{w}) \stackrel{\text{def}}{=} g'(\mathbf{w}^\mathsf{T}\mathbf{s})\mathbf{w} - g(\mathbf{w}^\mathsf{T}\mathbf{s})\mathbf{s}.$$

We denote by $\boldsymbol{h}_i$ the ith component of the random vector $\boldsymbol{h}(\mathbf{e}_1)$, and by $\nabla\boldsymbol{h}_{ij}$ the ijth component of the matrix $\nabla\boldsymbol{h}(\mathbf{e}_1)$. Now we are ready to announce the main result of the chapter:

Proposition 6.1.1. *We have*

$$\sqrt{N}\nabla\boldsymbol{f}(\mathbf{e}_1,\mu_N^1)\xrightarrow[N\to\infty]{\mathscr{D}}\mathcal{N}(0,\boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\Sigma}=\frac{1}{|H(\mathbf{e}_1,\mu)|^4}\mathrm{Cov}\Big(-\|\nabla\boldsymbol{h}(\mathbf{e}_1,\mu)\|\sum_{i=2}^{d}\boldsymbol{h}_i\,\mathrm{vec}(\mathbf{e}_i\mathbf{e}_1^{\mathsf{T}}) +|H(\mathbf{e}_1,\mu)|\sum_{i=2,j=1}^{d}\nabla\boldsymbol{h}_{ij}\,\mathrm{vec}(\mathbf{e}_i\mathbf{e}_j^{\mathsf{T}})\Big), \tag{6.1.1}$$

and we recall that

$$\begin{aligned}H(\mathbf{e}_1,\mu)&=\mathbb{E}[g'(s_1)-g(s_1)s_1],\\ \|\nabla\boldsymbol{h}(\mathbf{e}_1,\mu)\|&=|\mathbb{E}[s_1g''(s_1)+(1-s_1^2)g'(s_1)]|.\end{aligned}$$

Remark 6.1.2. The asymptotic covariance matrix $\boldsymbol{\Sigma}$ of $\nabla\boldsymbol{f}(\mathbf{e}_1,\mu_N^1)$ depends not only on the probability distribution of s_1, but also on the probability distribution of the rest source signals, according to (6.1.1).

In the 2-dimensional case, formula (6.1.1) can be significantly simplified. We can easily derive the following corollary:

Corollary 6.1.3. *In the case $d=2$, we have*

$$\mathrm{Trace}(\boldsymbol{\Sigma})=\frac{1}{|H(\mathbf{e}_1,\mu)|^4}\Big(\|H(\mathbf{e}_1,\mu)\|^2\mathrm{Cov}(\nabla\boldsymbol{h}_{22})+\mathrm{Cov}\big(|H(\mathbf{e}_1,\mu)|\nabla\boldsymbol{h}_{21} -\|\nabla\boldsymbol{h}(\mathbf{e}_1,\mu)\|\boldsymbol{h}_2\big)\Big),$$

where

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{h}_2)&=\mathbb{E}[g^2(s_1)]\\ \mathrm{Cov}(\boldsymbol{h}_2,\nabla\boldsymbol{h}_{21})&=\mathbb{E}[s_1g(s_1)g'(s_1)]\\ \mathrm{Cov}(\nabla\boldsymbol{h}_{21})&=\mathbb{E}[(s_1g'(s_1))^2]\\ \mathrm{Cov}(\nabla\boldsymbol{h}_{22})&=\mathbb{E}[(g'(s_1))^2]\mathbb{E}[s_2^4-1].\end{aligned}$$

6.2 Numerical results

Example 6.2.1. We consider the extraction of s_1 from a mixture of two source signals s_1, s_2 that have probability distributions including the uniform distribution, the Laplace distribution, the generalized Gaussian distribution

GG(α) and the Gaussian mixture distribution GM(p, m). Using the formula introduced in Corollary 6.1.3, we calculate Trace($\mathbf{\Sigma}$), for each nonlinearity functions including "kurtosis", "Gauss" and "tanh", and for each combination of probability distributions listed above. The optimal nonlinearity and the corresponding Trace($\mathbf{\Sigma}$) is displayed in Table 6.1 for each couple (s_1, s_2). From the table, we observe that the nonlinearity functions "kurtosis" and "Gauss" yield a faster convergence speed compared to "tanh".

($\mathbf{s}_1$, $\mathbf{s}_2$)	uniform	Laplace	GM($\frac{1}{6}$, 2)	GM($\frac{1}{10}$, 2.5)	GG(0.5)	GG(3)
uniform	(kurt, 9.00)	(gaus, 43.8)	(gaus, 28.9)	(gaus, 40.3)	(gaus, 105)	(gaus, 16.0)
Laplace	(kurt, 4.82)	(kurt, 30.1)	(kurt, 16.7)	(kurt, 26.9)	(kurt, 85.0)	(kurt, 8.55)
GM($\frac{1}{6}$, 2)	(gaus, 22.9)	(gaus, 85.3)	(gaus, 51.7)	(gaus, 76.6)	(gaus, 218)	(gaus, 31.7)
GM($\frac{1}{10}$, 2.5)	(gaus, 6.16)	(gaus, 31.4)	(gaus, 18.0)	(gaus, 28.2)	(gaus, 86)	(gaus, 9.89)
GG(0.5)	(kurt, 0.72)	(kurt, 4.52)	(kurt, 2.51)	(kurt, 4.04)	(kurt, 13.6)	(kurt, 1.28)
GG(3)	(kurt, 51.5)	(gaus, 295)	(kurt, 178)	(gaus, 270)	(gaus, 743)	(kurt, 91.3)

Table 6.1: The optimal choice of nonlinearity function and the corresponding asymptotic covariance matrix.

6.3 Proofs

6.3.1 Proof of Proposition 6.1.1

The proof of Proposition 6.1.1 relies on the following result, which is well known.

Lemma 6.3.1. *Let $\{\mathbf{y}_N\}$ be a sequence of $\mathbb{R}^m$-random vectors such that $\sqrt{N}(\mathbf{y}_N - \mathbf{y}) \xrightarrow[N\to\infty]{\mathscr{D}} \mathcal{N}(0, \mathbf{\Sigma})$ for some vector $\mathbf{y} \in \mathbb{R}^m$. Then for any mapping $\mathcal{P} : \mathbb{R}^m \to \mathbb{R}^n$ that has continuous second-order derivative, we have*

$$\sqrt{N}\Big(\mathcal{P}(\mathbf{y}_N) - \mathcal{P}(\mathbf{y})\Big) \xrightarrow[N\to\infty]{\mathscr{D}} \mathcal{N}\Big(0, \nabla\mathcal{P}(\mathbf{y})\mathbf{\Sigma}\nabla\mathcal{P}(\mathbf{y})^{\mathsf{T}}\Big). \tag{6.3.1}$$

Proof. By the second-order Taylor's formula, we get

$$\mathcal{P}(\mathbf{y}_N) = \mathcal{P}(\mathbf{y}) + \nabla\mathcal{P}(\mathbf{y})(\mathbf{y}_N - \mathbf{y}) + \mathcal{O}(\|\mathbf{y}_N - \mathbf{y}\|^2).$$

It follows that

$$\sqrt{N}\Big(\mathcal{P}(\mathbf{y}_N) - \mathcal{P}(\mathbf{y})\Big) = \nabla\mathcal{P}(\mathbf{y})\left\{\sqrt{N}\Big(\mathbf{y}_N - \mathbf{y}\Big)\right\} + \sqrt{N}\mathcal{O}(\|\mathbf{y}_N - \mathbf{y}\|^2). \tag{6.3.2}$$

On the one hand, from the hypothesis we deduce that

$$\nabla\mathcal{P}(\mathbf{y})\left\{\sqrt{N}\Big(\mathbf{y}_N - \mathbf{y}\Big)\right\} \xrightarrow[N\to\infty]{\mathscr{D}} \mathcal{N}\Big(0, \nabla\mathcal{P}(\mathbf{y})\mathbf{\Sigma}\nabla\mathcal{P}(\mathbf{y})^{\mathsf{T}}\Big). \tag{6.3.3}$$

On the other hand, $\|\mathbf{y}_N - \mathbf{y}\|$ is of order $\mathcal{O}_p(N^{-1/2})$. Hence we have

$$\sqrt{N}\mathcal{O}(\|\mathbf{y}_N - \mathbf{y}\|^2) \xrightarrow[N\to\infty]{\mathbb{P}} 0. \tag{6.3.4}$$

Letting $N \to \infty$ in equality (6.3.2) and noticing (6.3.3) and (6.3.4), we achieve (6.3.1). □

Proof of Proposition 6.1.1
The idea of the proof is quite straightforward. First, we notice that

$$\nabla \boldsymbol{f}(\mathbf{e}_1, \mu_N^1) = \frac{(\|\boldsymbol{h}(\mathbf{e}_1, \mu_N^1)\|^2 \mathbf{I} - \boldsymbol{h}(\mathbf{e}_1, \mu_N^1)\boldsymbol{h}^\mathsf{T}(\mathbf{e}_1, \mu_N^1))\nabla \boldsymbol{h}(\mathbf{e}_1, \mu_N^1)}{\|\boldsymbol{h}(\mathbf{e}_1, \mu_N^1)\|^3},$$

which is a function of $\boldsymbol{h}(\mathbf{e}_1, \mu_N^1)$ and $\nabla \boldsymbol{h}(\mathbf{e}_1, \mu_N^1)$. By the central limit theorem, we have

$$\sqrt{N}\left\{\mathrm{vec}\Big(\boldsymbol{h}(\mathbf{e}_1, \mu_N^1), \nabla \boldsymbol{h}(\mathbf{e}_1, \mu_N^1)\Big) - \mathrm{vec}\Big(\boldsymbol{h}(\mathbf{e}_1, \mu), \nabla \boldsymbol{h}(\mathbf{e}_1, \mu)\Big)\right\} \tag{6.3.5}$$

$$\xrightarrow[N\to\infty]{\mathscr{D}} \mathcal{N}\Big(0, \mathrm{Cov}\Big(\mathrm{vec}(\boldsymbol{h}(\mathbf{e}_1), \nabla \boldsymbol{h}(\mathbf{e}_1))\Big)\Big), \tag{6.3.6}$$

where $\mathrm{vec}(\mathbf{w}, \mathbf{M})$ denotes the operation that combines $\mathbf{w}$ and $\mathrm{vec}(\mathbf{M})$ into one long column vector for any vector $\mathbf{w}$ and matrix $\mathbf{M}$. Besides, due to the fact

$$\|\boldsymbol{h}(\mathbf{e}_1, \mu_N^1)\|^3 \xrightarrow[N\to\infty]{a.s.} \|\boldsymbol{h}(\mathbf{e}_1, \mu)\|^3 \neq 0, \tag{6.3.7}$$

we need only to study the numerator of $\nabla \boldsymbol{f}(\mathbf{e}_1, \mu_N^1)$ thanks to Slutsky's Lemma. Applying Lemma 6.3.1, we are going to achieve the task as following.

Let's define the mapping $\mathcal{P} : \mathbb{R}^d \times \mathbb{R}^{d\times d} \mapsto \mathbb{R}^{d\times d}$ by

$$\mathcal{P}(\mathbf{w}, \mathbf{M}) \stackrel{\text{def}}{=} (\|\mathbf{w}\|^2 \mathbf{I} - \mathbf{w}\mathbf{w}^\mathsf{T})\mathbf{M}.$$

We have $\mathcal{P}\Big(\boldsymbol{h}(\mathbf{e}_1, \mu), \nabla \boldsymbol{h}(\mathbf{e}_1, \mu)\Big) = 0$ and

$$\nabla \boldsymbol{f}(\mathbf{e}_1, \nu) = \frac{\mathcal{P}\Big(\boldsymbol{h}(\mathbf{e}_1, \nu), \nabla \boldsymbol{h}(\mathbf{e}_1, \nu)\Big)}{\|\boldsymbol{h}(\mathbf{e}_1, \nu)\|^3} \tag{6.3.8}$$

for any measure ν. Thus we need only to study the limit of

$$\sqrt{N}\mathcal{P}\Big(\boldsymbol{h}(\mathbf{e}_1, \mu_N^1), \nabla \boldsymbol{h}(\mathbf{e}_1, \mu_N^1)\Big)$$

as N tends to infinity.

Let us denote respectively by $\nabla_{\mathbf{w}_i}\mathcal{P}(\cdot,\cdot)$ and $\nabla_{\mathbf{M}_{ij}}\mathcal{P}(\cdot,\cdot)$ the partial derivative of $\mathcal{P}(\cdot,\cdot)$ with respect to the ith component of its first argument, and the ijth component of its second argument. For $1 \leq i,j \leq d$ We have:

$$\begin{aligned}\nabla_{\mathbf{w}_i}\mathcal{P}(\mathbf{w},\mathbf{M}) &= 2\mathbf{w}_i\mathbf{M} - (\mathbf{e}_i\mathbf{w}^\mathsf{T} + \mathbf{w}\mathbf{e}_i^\mathsf{T})\mathbf{M},\\ \nabla_{\mathbf{M}_{ij}}\mathcal{P}(\mathbf{w},\mathbf{M}) &= (\|\mathbf{w}\|^2\mathbf{e}_i\mathbf{e}_j^\mathsf{T} - \mathbf{w}\mathbf{w}^\mathsf{T}\mathbf{e}_i\mathbf{e}_j^\mathsf{T}).\end{aligned}$$

Now, we are going to calculate the first order partial derivatives of $\mathcal{P}(\cdot,\cdot)$ at the point

$$\widetilde{\mathbf{H}} \overset{\text{def}}{=} \Big(\boldsymbol{h}(\mathbf{e}_1,\mu), \nabla\boldsymbol{h}(\mathbf{e}_1,\mu)\Big) \in \mathbb{R}^d \times \mathbb{R}^{d\times d}.$$

Since

$$\begin{aligned}\boldsymbol{h}(\mathbf{e}_1,\mu) &= \mathbb{E}[g'(s_1) - s_1 g(s_1)]\mathbf{e}_1,\\ \nabla\boldsymbol{h}(\mathbf{e}_1,\mu) &= \mathbb{E}[s_1 g''(s_1) + (1-s_1^2)g'(s_1)]\mathbf{e}_1\mathbf{e}_1^\mathsf{T},\end{aligned}$$

it follows that

$$\nabla_{\mathbf{w}_1}\mathcal{P}\big(\widetilde{\mathbf{H}}\big) = 0 \tag{6.3.9}$$

$$\nabla_{\mathbf{w}_i}\mathcal{P}\big(\widetilde{\mathbf{H}}\big) = -\|\boldsymbol{h}(\mathbf{e}_1,\mu)\|\|\nabla\boldsymbol{h}(\mathbf{e}_1,\mu)\|\mathbf{e}_i\mathbf{e}_1^\mathsf{T} \tag{6.3.10}$$

$$\nabla_{\mathbf{M}_{1j}}\mathcal{P}\big(\widetilde{\mathbf{H}}\big) = 0 \tag{6.3.11}$$

$$\nabla_{\mathbf{M}_{ij}}\mathcal{P}\big(\widetilde{\mathbf{H}}\big) = \|\boldsymbol{h}(\mathbf{e}_1,\mu)\|^2\mathbf{e}_i\mathbf{e}_j^\mathsf{T} \tag{6.3.12}$$

for $i = 2,...,d, j = 1,...,d$.

Applying Lemma 6.3.1, we obtain

$$\begin{aligned}&\sqrt{N}\left\{\mathcal{P}\Big(\boldsymbol{h}(\mathbf{e}_1,\mu_N^1), \nabla\boldsymbol{h}(\mathbf{e}_1,\mu_N^1)\Big) - \mathcal{P}\big(\widetilde{\mathbf{H}}\big)\right\}\\ &\quad= \sqrt{N}\mathcal{P}\Big(\boldsymbol{h}(\mathbf{e}_1,\mu_N^1), \nabla\boldsymbol{h}(\mathbf{e}_1,\mu_N^1)\Big) \xrightarrow[N\to\infty]{\mathscr{D}} \mathcal{N}(0,\boldsymbol{\Sigma}_0),\end{aligned} \tag{6.3.13}$$

where

$$\begin{aligned}\boldsymbol{\Sigma}_0 &= \nabla\mathcal{P}(\widetilde{\mathbf{H}})\,\mathrm{Cov}\Big(\mathrm{vec}(\boldsymbol{h}(\mathbf{e}_1), \nabla\boldsymbol{h}(\mathbf{e}_1))\Big)\,\nabla\mathcal{P}(\widetilde{\mathbf{H}})^\mathsf{T}\\ &= \mathrm{Cov}\left(\sum_{i=1}^d \boldsymbol{h}_i\mathrm{vec}\Big(\nabla_{\mathbf{w}_i}\mathcal{P}\big(\widetilde{\mathbf{H}}\big)\Big) + \sum_{i,j=1}^d \nabla\boldsymbol{h}_{ij}\mathrm{vec}\Big(\nabla_{\mathbf{M}_{ij}}\mathcal{P}\big(\widetilde{\mathbf{H}}\big)\Big)\right)\\ &= \mathrm{Cov}\Big(-\|\nabla\boldsymbol{h}(\mathbf{e}_1,\mu)\|\|\boldsymbol{h}(\mathbf{e}_1,\mu)\|\sum_{i=2}^d \boldsymbol{h}_i\mathrm{vec}(\mathbf{e}_i\mathbf{e}_1^\mathsf{T})\\ &\qquad +\|\boldsymbol{h}(\mathbf{e}_1,\mu)\|^2\sum_{i=2,j=1}^d \nabla\boldsymbol{h}_{ij}\mathrm{vec}(\mathbf{e}_i\mathbf{e}_j^\mathsf{T})\Big),\end{aligned}$$

thanks to (6.3.9)-(6.3.12). Finally, we deduce from (6.3.7), (6.3.8) and (6.3.13) that

$$\sqrt{N}\nabla\boldsymbol{f}(\mathbf{a},\mu_N^1) \xrightarrow[N\to\infty]{\mathscr{D}} \mathcal{N}\Big(0, \frac{\boldsymbol{\Sigma}_0}{\|\boldsymbol{h}(\mathbf{e}_1,\mu)\|^6}\Big),$$

from which the result follows.

6.3.2 Proof of Corollary 6.1.3

In the case of $d = 2$, we need only to compute the covariance matrix of $(\boldsymbol{h}_2, \nabla\boldsymbol{h}_{21}, \nabla\boldsymbol{h}_{22})$. We recall that

$$\begin{aligned} \boldsymbol{h}_2 &= -s_2 g(s_1) \\ \mathrm{Cov}(\boldsymbol{h}_2) &= \mathbb{E}[g^2(s_1)]. \end{aligned}$$

Concerning the matrix $\nabla\boldsymbol{h}(\mathbf{e}_1)$, we have

$$\begin{aligned} \nabla\boldsymbol{h}_{22} &= g'(s_1)(1 - s_2^2) \\ \nabla\boldsymbol{h}_{21} &= -s_2 s_1 g'(s_1). \end{aligned}$$

Hence, the covariance matrix of the vector $(\boldsymbol{h}_2, \nabla\boldsymbol{h}_{21}, \nabla\boldsymbol{h}_{22})$ is given by:

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{h}_2) &= \mathbb{E}[g^2(s_1)] \\ \mathrm{Cov}(\boldsymbol{h}_2, \nabla\boldsymbol{h}_{21}) &= \mathbb{E}[s_1 g(s_1) g'(s_1)] \\ \mathrm{Cov}(\nabla\boldsymbol{h}_{21}) &= \mathbb{E}[(s_1 g'(s_1))^2] \\ \mathrm{Cov}(\nabla\boldsymbol{h}_{22}) &= \mathbb{E}[(g'(s_1))^2]\mathbb{E}[s_2^4 - 1]. \end{aligned}$$

which achieves the proof.

Chapter 7

Conclusion and Perspective

7.1 Summary of results

In this thesis, we have given a unified study of both theoretical and empirical versions of the deflation based (one-unit) FastICA algorithm,

In Chapter 3, we proved that the theoretical FastICA algorithm converges to a column of the mixing matrix with at least a quadratic convergence speed. In particular, if the underlying nonlinearity function is kurtosis, the convergence speed is even cubic. Although these results (e.g. Theorem 3.3.4, Proposition 3.3.5) are already well-known, our approach is novel and rigorous. We calculated the gradient of the FastICA function, and showed that the gradient would vanish at the columns of the mixing matrix $\mathbf{A}$. From this fact, the quadratic convergence speed of the theoretical FastICA algorithm is immediately derived. In the case of kurtosis nonlinearity, the second derivative of the FastICA function vanishes, leading to a cubic convergence speed. Moreover, we proved that any local minimizer of the theoretical contrast function is always a fixed point of the theoretical FastICA function. This characterization is, to our knowledge, new. It also gave some insight to the study of empirical FastICA algorithm.

In Chapter 4, we studied the convergence of four empirical FastICA algorithms. The main result of this chapter is that each empirical FastICA algorithm converges to a local minimizer of the respective empirical contrast function with probability one, provided that the sample size is large enough. As a corollary of this result, we showed that each FastICA estimator, defined as the limit of the empirical FastICA algorithm, is a consistent estimator of a column of the mixing matrix. Before our attempt, the convergence of FastICA algorithm in the finite sample case was only supported by numerical simulation but never theoretically confirmed.

In Chapter 5, we gave the explicit closed form of the asymptotic covariance matrix of our FastICA estimators. Although similar results already exist in the literature, our approach is the only mathematically rigourous one, and we addressed four different scenarios all together, which enables the measuring of the effect of data centering and whitening. We also made a comparison of the existing results was made in Section 5.1.1.

In Chapter 6, we studied the gradient of the empirical FastICA algorithm and derived a new criterion of optimality of the choice of nonlinearity function. This new criterion consists of optimizing the asymptotic covariance

matrix of the gradient of the empirical FastICA function at the column of the mixing matrix, such that one can attain the fastest possible convergence speed of the empirical FastICA algorithm. We derived the explicit form of the latter asymptotic covariance matrix and gave some numerical example for two sources.

7.2 Upcoming challenges

Although this thesis gives elements of answer to some interesting issues concerning the empirical FastICA algorithm, there are still questions remaining unsolved.

7.2.1 Spurious local optima

Spurious local optimizers of the contrast function are those optimizers that do not correspond to any demixing vectors. Convergence of ICA algorithm to spurious local optimizers are the very thing we hope to avoid. There are two ways to achieve this: we can either choose carefully the contrast function so that it does not possess spurious local optima (this is the case for mutual information based contrast function, see (Comon, 1994)), or we design an ICA algorithm which can somehow, intelligently converge to the right optima. In the continued study we actually proved the following inclusion for FastICA:

$$\mathbb{D} \subset \mathbb{L} \subset \mathbb{O} \subset \mathbb{F},$$

where

$\mathbb{D} \stackrel{\text{def}}{=} \{\pm \mathbf{a}_i,\ i = 1, \ldots, d\}$;

$\mathbb{F} \stackrel{\text{def}}{=} \{\mathbf{v} \in \mathbb{R}^d :\ \boldsymbol{f}(\mathbf{v}, \mu) = \pm \mathbf{v}\}$;

$\mathbb{L} \stackrel{\text{def}}{=} \{\mathbf{v} \in \mathbb{F} :\ \|\nabla \boldsymbol{f}(\mathbf{v}, \mu)\| < 1\}$;

$\mathbb{O} \stackrel{\text{def}}{=} \{\mathbf{v} \in \mathcal{S} :\ \mathbf{v}$ is a local optimizer of $G(\cdot, \mu)\}$.

This inclusion suggests that unlike the other gradient-ascent type algorithms, FastICA algorithm does not searches blindly all the local optima of the contrast function; it can automatically filter those spurious optimizers which are not stable fixed points. Now the vital question is *if any of these inclusions, especially* $\mathbb{D} \subset \mathbb{L}$*, is indeed an equality.* If we can show, or at least give some sufficient condition under which $\mathbb{D} = \mathbb{L}$, i.e. FastICA always yield demixing vectors, then we can assert that the algorithm is theoretically reliable. For kurtosis nonlinearity, this is done in (Douglas, 2003). Nevertheless, we believe that the equality $\mathbb{D} = \mathbb{L}$ cannot be true in general, although we have never encountered systematic violation of it in repeated experiments. We are not able to provide a proof or come up with a counter-example at the moment.

At least, it is easy to give an example where the inclusion $\mathbb{L} \subset \mathbb{O}$ is strict. It is known that in the simplest 2-dimensional case, if the two source signal have the same distribution, then there must exist local optima that do not correspond to the demixing vectors (see e.g. Fig. 2.4). This is the case where we have

$$\mathbb{D} = \mathbb{L} \subset \mathbb{O} = \mathbb{F}.$$

7.2.2 Convergence radius

In Section 3.4.2, we have briefly talked about the convergence radius of the FastICA algorithm. For each source signal s_i, its convergence radius is defined to be the largest real number r, such that if the initial input $\mathbf{w}_0$ of the FastICA algorithm lies in the ball $\mathscr{B}_r(\mathbf{a}_i)$, then the FastICA algorithm is guaranteed to converge to the corresponding demixing vector $\mathbf{a}_i$. Experiments shows that source signals that are comparatively closer to Gaussian correspond to a smaller convergence radius. We believe that this can be theoretically proved, with the underlying measure of Gaussianity:

$$|H(\mathbf{a}_i, \mu)| = |\mathbb{E}[g'(s_i) - g(s_i)s_i]|.$$

This quantity vanishes if s_i has a Gaussian distribution, as is shown in Section 3.1. We predict that if the source signal s_i, non-necessarily being Gaussian, has a very small $|\mathbb{E}[g'(s_i) - g(s_i)s_i]|$, then it would be very difficult to recover with one-unit FastICA and the underlying nonlinearity $G(\cdot)$, since most initial iterate of the algorithm would make the algorithm converge to elsewhere. Using another nonlinearity may resolve this problem though.

7.2.3 Convergence and asymptotic behavior of FastICA for the extraction of several sources

In this thesis, we have only considered the extraction of only one source using one-unit FastICA. This scenario is easy to analyze, since the deflation procedure (2.1.23) can be omitted. In this case, if we are to recover all the sources, the starting points need to locate in the neighborhood of different columns of the mixing matrix, so that the algorithm can converge to different limits. However, in real world problem, this may be unrealistic or computational costly. For example, many different starting point on $\mathcal{S}$ may end up yielding the same source, while some source (especially those whose probability distribution is close to Gaussian) may never be recovered due to small convergence radius, as is previously explained In order to effectively extract all the sources, one need to implement either symmetrical FastICA, or the original one-unit FastICA with a decorrelation procedure: to avoid recovering twice a same source, at each iteration step, the i-th recovered source must be orthogonal to the $i-1$ previously extracted sources.

The questions that motivate this thesis remain still open for these versions of FastICA. Namely, when extracting the 2nd, 3rd, ..., n-th sources, how does the additional procedure influence the convergence of the one-unit FastICA algorithm? What is the asymptotic covariance matrix of the sequentially obtained FastICA estimators? What about the symmetrical version of FastICA in the finite sample case? The difficulty lies in the fact that if the sample size is finite, the minimizers of the empirical contrast function does not form an orthogonal matrix in general, while both one-unit FastICA and symmetrical FastICA can only yield an orthogonal estimate of the demixing matrix. This means that unlike the case of extracting only one source, when trying to recover all the sources, the FastICA estimators cannot be the local minimizers of the empirical contrast function. This fact is problematic since the whole analysis presented in this thesis is based upon the link between the local minimizers of the contrast function and fixed points of the FastICA function.

By generalizing some ideas used in this thesis, we derived heuristically the asymptotic covariance matrix for symmetrical FastICA:

$$\mathbf{R}_i = \sum_{j\neq i}^{d} \frac{\beta_i - \alpha_i^2 + \beta_j - \alpha_j^2 + (\rho_j - \alpha_j)^2 - \eta_j^2}{(|\rho_i - \alpha_i| + |\rho_j - \alpha_j|)^2} \mathbf{b}_j \mathbf{b}_j^{\mathsf{T}} + \kappa_i \mathbf{a}_i \mathbf{a}_i^{\mathsf{T}} + \Big(\sum_{j\neq i}^{d} \frac{\operatorname{sign}(\rho_j - \alpha_j)\eta_j \mathbf{a}_j}{|\rho_i - \alpha_i| + |\rho_j - \alpha_j|} \Big) \Big(\sum_{j\neq i}^{d} \frac{\operatorname{sign}(\rho_j - \alpha_j)\eta_j \mathbf{a}_j^{\mathsf{T}}}{|\rho_i - \alpha_i| + |\rho_j - \alpha_j|} \Big).$$

This formula should be correct since it coincides with the formula given in (Tichavsky et al., 2006) using another heuristic method. However, some issues are still to be resolved if we hope to render the whole argument rigorous.

References

Andrews, D. W. K. (1992, June). Generic uniform convergence. *Econometric Theory*, *02*(9). (Cited on page 49.)

Bierens, H. (2005). Introduction to the mathematical and statistical foundations of econometrics. In (p. 171). Cambridge University Press. (Cited on page 49.)

Brandstein, M., & Ward, D. B. (2001). *Microphone arrays: Signal processing techniques and applications*. New York: Springer. (Cited on page 3.)

Cardoso, J. F., & Souloumiac, A. (1993, December). Blind beamforming for non-Gaussian signals. *IEEE Proceedings-F*, *140*(6), 362-370. (Cited on page 4.)

Chevalier, P., Albera, L., Comon, P., & Ferreol, A. (2004, July). Comparative performance analysis of eight blind source separation methods on

radio-communications signals. In *Joint conf. neural netw.* (p. 273-278). (Cited on page 4.)

Comon, P. (1994, April). Independent component analysis: a new concept? *Signal Processing*, *36*(3), 287-314. (Cited on pages 3, 8, 9, 11, 13, 14 and 88.)

Comon, P., & Jutten, C. (2010). Handbook of blind source separation: Independent component analysis and applications. In (p. 179-227). Academic Press. (Cited on pages 3 and 11.)

Comon, P., & Moreau, E. (1997, April). Improved contrast dedicated to blind separation in communications. In *Proc. icassp-97.* (Cited on page 4.)

Delfosse, N., & Loubaton, P. (1995, July). Adaptive blind separation of independent sources. *Signal Processing*, *45*, 59-83. (Cited on pages 4 and 21.)

Dermoune, A., Rahmania, N., & Wei, T. (2012). General linear mixed model and signal extraction problem with constraint. *Journal of Multivariate Analysis*, *105*, 311-321. (Cited on page 3.)

Dermoune, A., & Wei, T. (2013, April). FastICA algorithm: Five criteria for the optimal choice of the nonlinearity function. *IEEE transaction on Signal Processing*, *61*(8), 2078-2087. (Cited on page 3.)

Donoho, D. (1981). On minimum entropy deconvolution. In *Applied time series analysis ii* (p. 565-609). New York: Academic. (Cited on page 16.)

Douglas, S. (2003, apr). On the convergence behavior of the fastica algorithm. In *Proc. 4th symp. independent component analysis blind source separation* (p. 409-414). (Cited on pages 4 and 88.)

Hyvärinen, A. (1997). One-unit contrast functions for independent component analysis: A statistical analysis. In *Proc. ieee nnsp workshop '97.* (Cited on pages 4 and 65.)

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, *10*(3), 626-634. (Cited on pages 4, 15, 16, 17, 20, 22, 28, 34 and 35.)

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis.* New York: Wiley-Interscience. (Cited on pages 3 and 8.)

Hyvärinen, A., & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, *9*(7), 1483-1492. (Cited on pages 4 and 34.)

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, *13*(4-5), 411-430. (Cited on pages 3, 4, 8, 13, 14 and 15.)

Jutten, C. (1987). *Calcul neuromimétique et traitement du signal: analyse en composantes indpendantes.* Thèse détat es sciences physiques, UJF-INP Grenoble. (Cited on page 3.)

Jutten, C., & Comon, P. (n.d.). *De la séparation des sources à l'analyse en composantes indépendante.* (Cited on page 3.)

Jutten, C., & Taleb, A. (2000). Source separation: From dusk till dawn. In *2nd int. workshop on independent component analysis and blind source separation (ica 2000).* (Cited on page 3.)

Leshem, A., & van der Veen, A.-J. (2008, September). Blind source separation: the location of local minima in the case of finitely many samples. *IEEE transactions on Signal Processing*, *56*(9), 4340-4353. (Cited on page 4.)

Luenberger, D. G., & Ye, Y. (2008). Linear and nonlinear programming. In (Third ed.). Springer. (Cited on page 67.)

Makeig, S., Bell, A., Jung, T.-P., & Sejnowski, T. (1996). Independent component analysis of electroen-cephalographic data. *Advances in Neural Information Processing Systems*(9), 145-151. (Cited on page 3.)

Makino, S., Lee, T.-W., & Sawada, H. (2007). *Blind speech separation.* Dordrecht, the Netherlands: Springer. (Cited on page 3.)

M. Ichir, A., M. Mohammad-Djafari. (2006, July). Hidden Markov models for wavelet-based blind source separation. *IEEE transactions on Image Processing*, *15*(7), 1887-1899. (Cited on page 3.)

Nesterov, Y. (2004). *Introductory lectures on convex optimization: A basic course.* Boston/Dordrecht/London: Kluwer. (Cited on page 67.)

Newey, W. K. (1991, July). Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, *59*(4), 1161-1167. (Cited on page 49.)

Oja, E. (2002, November). Convergence of the symmetrical FastICA algorithm. In *9th int. conf. neural information processing (iconip).* (Cited on pages 4, 17 and 35.)

Oja, E., & Yuan, Z. (2006). The FastICA algorithm revisited: Convergence analysis. *IEEE transactions on Neural Networks*, *17*(6). (Cited on pages 4, 17 and 51.)

Ollila, E. (2010, March). The deflation-based FastICA estimator: Statistical analysis revisited. *IEEE transactions on Signal Processing*, *58*(3). (Cited on pages 4, 65 and 66.)

Regalia, P. A., & Kofidis, E. (2003, July). Monotonic convergence of fixed-point algorithms for ICA. *IEEE transactions on Neural Network*, *14*(4), 943-949. (Cited on pages 4, 22 and 37.)

Reyhani, N., Ylipaavalniemi, J., Vigario, R., & Oja, E. (2012). Consistency and asymptotic normality of FastICA and bootstrap FastICA. *Signal Processing*, *92*, 1767-1778. (Cited on pages 65 and 66.)

Shen, H., Kleinsteuber, M., & Hüper, K. (2008, June). Local convergence analysis of FastICA and related algorithms. *IEEE transactions on Neural Network*, *19*(6), 1022-1032. (Cited on pages 22 and 28.)

Shimizu, A., Hyvärinen, A., Yutaka, K., Hoyer, P., & Kerminen, A. J. (2006). Testing signifcance of mixing and demixing coefficients in ICA. In *Int. conf. independent component analysis (ica 2006).* (Cited on page 65.)

Stone, J. V. (2004, sep). Independent component analysis: A tutorial introduction. A Bradfort Book. (Cited on page 3.)

Tichavsky, P., Koldovsky, Z., & Oja, E. (2006, April). Performance analysis of the FastICA algorithm and Cramer-Rao bounds for linear independent component analysis. *IEEE transactions on Signal Processing*, *54*(4), 1189-1203. (Cited on pages 4, 40, 51, 65, 66 and 90.)

Tugnait, J. K. (1997, March). Identification and deconvolution of multichannel non-Gaussian processes using higher order statistics and inverse filter criteria. *IEEE Transactions on Signal Processing*, *45*, 658-672. (Cited on page 4.)

van der Vaart, A. (2000). Asymptotic statistics. In (chap. 5). Cambridge University Press. (Cited on page 72.)

Vigario, R., & Oja, E. (2008). BSS and ICA in neuroinformatics: From current practices to open challenges. *IEEE Reviews in Biomedical Engineering*(1), 50-61. (Cited on page 3.)

Vrins, F. D. (2007). *Contrast properties of entropic criteria for blind source separation, a unifying framework based on information-theoretic inequalities*. Unpublished doctoral dissertation, Université Catholique de Louvain. (Cited on pages 11, 14 and 35.)

Waheed, K., & Salam, F. M. (2002, August). Blind source recovery using an adaptive generalized Gaussian score function. In *in proc. 45th midwest symp. circuits systems (mwscas)* (p. 656-659). (Cited on page 40.)

Zarzoso, V., & Comon, P. (2010, February). Robust independent component analysis by iterative maximization of the kurtosis contrast with algebraic optimal step size. *IEEE Transactions on Neural Networks*, *21*(2), 248-261. (Cited on page 4.)